ARCHETYPAL ANALYSIS FOR INTERVAL DATA IN MARKETING RESEARCH[1]

Maria Rosaria D'Esposito

Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte don Melillo, 84084 Fisciano (ITALY)
mdesposito@unisa.it

Francesco Palumbo

Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata, Via Crescimbeni, 20, 62100, Macerata (ITALY)
francesco.palumbo@unimc.it

Giancarlo Ragozini

Dipartimento di Sociologia "G. Germani", Università Federico II, Napoli Vico Monte della Pietà 1, 80138 Napoli (ITALY)
giragoz@unina.it

Abstract

A typical problem in marketing research consists of segmenting and clustering products and/or consumers. However, classical cluster analysis and segmentation may fail in the interpretability as they tend to identify average consumers or products, sometimes not well-separated. In this framework, archetypal analysis has been introduced to find extreme segments and well separated typical consumers. On the other hand, we notice that often product attributes and consumer preferences could be more adequately expressed by a range of values in which attributes/preferences may vary. To face these two issues, in this work, we propose an extension of archetypal analysis to the case of interval data, providing a definition of archetypes using the Hausdorff distance, analizing their geometric properties, and offering some appropriate visualization tools. We present also an illustrative example on preference data.

Keywords: Archetypal Consumer; Hausdorff Distance; Market Segmentation.

---

[1] *This paper was financially supported by PRIN2003 grant: "Multivariate Statistical and Visualization Methods to Analyze, Summarize, and Evaluate Performance Indicators"*

1. INTRODUCTION

A typical problem in marketing research consists of segmenting and clustering products and/or consumers. In this framework, some issues are the study of the heterogeneity of consumer behaviors and of the different combinations of product attributes which may determine consumer preferences. Classical multivariate statistical methods are largely applied for these aims. More recently, in an attempt to provide complete market segmentation more sophisticated techniques, that rely on the selection of a response variable, have been exploited, such as latent class techniques, tree based segmentation like CHAID and CART (Ratner, 2003; De Sarbo *et al*., 2004).

However, two different issues arise. First, classical cluster analysis and segmentation may fail in the interpretability as they tend to identify average consumers or products, sometimes not well-separated. Secondly, rather than by a single value, often product attributes and consumer preferences could be more adequately expressed by a range of values in which attributes/preferences may vary.

To face the first problem, in the marketing literature, archetypal analysis (Cutler and Breiman, 1994) has been introduced to find extreme segments and well separated typical consumers (Riedesel, 2003; Elder and Pinnell 2003; Li *et al*., 2003). While, for the second issue, the notion of interval data may yield more adequate statistical methods that take into account the uncertainty of the preferences and the variability of product attributes.

In this work, we explore the possibility of extending the proposal of using archetypal analysis in marketing research to the case of interval data. The original proposal consists of considering archetypal analysis as a method for selecting some consumers/products able to clearly represent a market segmentation. In this paper we define the archetypes when consumer preferences related to different products or many products attributes are measured as intervals, we discuss the related computational issues, and we present exploratory tools to visually analyze them based on properties of both archetypes and interval data. In particular we propose to use some iconic plots such as the stars for interval data, and principal component analysis for interval data to inspect the archetype's composition and to compare observed consumer or products among them and with respect to the archetypes.

The paper is organized as follows: we present basic archetypal analysis, in-

troducing a computational geometry point of view in Section 2, while some notion on interval data are in Section 3. The definition of the archetypes for interval data, their computational issues and their graphical representation are in Section 4. An illustrative example on preferences data will be used along the whole paper. Some concluding remarks follow.

2. ARCHETYPAL ANALYSIS IN MARKETING RESEARCH

Marketing researches usually aim at defining subsets, segments of consumers or of products to gain insight into the market and consumer behavior, in order to design appropriate communication strategies. To understand and analyze the heterogeneity of consumer behaviors and/or of product feature combinations in a specific market, classical multivariate statistical methods, such as hierarchical and not hierarchical clustering, have been used (Wedel and Wagner, 2000). However, the classical segmentation techniques implicitly assume that there are several average consumers and try to find such average objects, i.e. the cluster centroids. This often yields to identify segments not clearly separated on important profiling variables (Morris and Schmolze, 2006). Moreover, some marketers have noticed that, at least in the world of marketing, very few people aspire "to be average" (Elder and Pinell, 2003). Finally, if the interest is on new products, the focus is not on average consumers, but on *switchers*, i.e. customers having extreme consumption behavior (Allenbey and Ginter, 1995).

All these arguments have led several marketers to introduce the idea of consumer archetypes (Morris and Schmolze, 2006), intended as consumer with extreme profiles, and to propose the use of archetypal analysis in marketing researches (Riedesel, 2003; Elder and Pinnell 2003; Li *et al*., 2003, Anderson and Weiner, 2004). Archetypal analysis (Cutler and Breiman, 1994) is a statistical method aiming at synthesizing a set of multivariate observations through few points not necessarily observed. These points, the archetypes, can be considered a sort of "pure" types as all the data points must be a mixture of them. In addition, to ensure that these "pure" points are as close as possible to the observed data, archetypes must be also a convex combination of the data points.

First applications of archetypal analysis have been in spatio-temporal dynamics and cellular flames (Stone and Cutler, 1996; Stone, 2002), in medicine and in astronomy (Chan *et al*., 2003). In performance analysis archetypes have been used for ordering multivariate performances (D'Esposito and Ragozini, 2004) and for benchmarking (Porzio et al, 2006).

Considering marketing research, archetypal analysis has been introduced to

find some "pure" consumer or product profiles lying on the edges of data that best exemplify the differences the segmentation is attempting to define. All other consumers or products are expressed as a probabilistic mixture of such extremes, and the convex combination coefficients are used to define a fuzzy segmentation (Elder and Pinnel, 2003).

Formally, the archetypes $\mathbf{a}_j$, $j = 1,\ldots,m$, should be those points in the p-dimensional Euclidean space such that:

$$\mathbf{x}_i' = \alpha_i'\mathbf{A} \tag{1}$$

with

$$\alpha_{ij} \geq 0 \quad \forall i,j \qquad \alpha_i'\mathbf{1} = 1 \quad \forall i, \tag{2}$$

where $\mathbf{x}_i'$, $i = 1,\ldots,n$, are the observed data, $\mathbf{A}$ is the archetype matrix with $\mathbf{a}_j'$ its j-th row, and α_i' is the vector of the convex combination coefficients of the m archetypes for the i-th data point, with generic elements α_{ij}, $j = 1,\ldots,m$.

At the same time, all the archetypes should be also a mixture of the observed data:

$$\mathbf{a}_j' = \beta_j'\mathbf{X} \tag{3}$$

with

$$\beta_{ji} \geq 0 \quad \forall j,i \qquad \beta_j'\mathbf{1} = 1 \quad \forall j, \tag{4}$$

where $\mathbf{X}$ is the observed data matrix, and the convex combination coefficient β_{ji}'s are the n elements of the β_j' vectors, i.e. the weights of the n observations in determining the j-th archetype.

By definition of convex hull, eqn.s (1) and (2) imply that all the data belong to the convex hull of the archetypes, that is the archetypes could be the vertices of any convex p-polytope including the data scatter. On the other hand, eqn.s (3) and (4) imply that archetypes belong to the convex hull of the data. Consequently, archetypes are the vertices of the data convex hull.

However, in practice, the number of the data convex hull vertices is generally too large to synthesize data through few pure types. For this reason, looking for a smaller number of pure types, and wishing to preserve their closeness to the data (eqn.s 3 and 4), Cutler and Breiman (1994) defined the archetypes as those m points that fulfill as much as possible eqn. (1), satisfying at the same time eqn.s (2), (3) and (4).

More precisely, let us rewrite eqn. (1) as $\mathbf{x}_i' - \alpha_i'\mathbf{A} = \mathbf{0}$. For the discussion above, if the number of archetypes is less than the number of the data convex hull

vertices, then eqn (1) does not hold. In particular, for the points i^* lying outside the convex hull of the archetypes, we have that $\|\mathbf{x}'_{i*} - \alpha'_{i*}\mathbf{A}\| > 0$, where $\|\cdot\|$ is the L_2 norm of a vector. The archetypes, given m, have been then defined as the points $(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ minimizing

$$\sum_{i=1}^{n} \left\|\mathbf{x}'_i - \alpha'_i\mathbf{A}\right\|, \tag{5}$$

holding equations (2),(3) and (4).

The solution to this minimization problem depends on m, and solutions are not nested as m varies. That is, denoting with $\mathbf{a}'_j(m)$ the j-th archetype for a given m, $\mathbf{a}'_j(m) \neq \mathbf{a}'_j(l)$, with $m \neq l$.

As for the choice of m, Cutler and Breiman (1994) suggest to look at the quantity:

$$RSS(m) = \sum_{i=1}^{n} \left\|\mathbf{x}'_i - \tilde{\mathbf{x}}'_i(m)\right\| \tag{6}$$

where $\tilde{\mathbf{x}}'_i(m) = \alpha'_i(m) \cdot \mathbf{A}(m)$ are the best approximations of the observations $\mathbf{x}'_i$ through the m archetypes. The residual sum of squares $RSS(m)$ is then the sum of the euclidean distances of the observed data from their best approximation, and therefore it measures to what extent the m archetypes synthesize the data. It is worth noting that, given the number of archetypes m, eqn. (5) is equivalent to eqn. (6), and hence minimizing the first equation is equivalent to minimize the second one.

Before advancing our proposal of extending the archetypal analysis to interval data in marketing researches, we first introduce in the next section the main concept related to interval data.

3. A BRIEF INTRODUCTION TO INTERVAL DATA

Given a set of statistical units, each unit is generally coded into an order p row vector, where p indicates the number of observed features measured by a single value.

Interval data represent a different way of coding variables, with respect to the classical *single-valued* data. There are, in fact, different sources of *incertitude* in the data suggesting the interval data coding: measurement errors, repeated measures, data usually reported in terms of *min* and *max* values (*e.g.* temperatures), instead of one single central tendency measure.

To exemplify, let us consider the *Juices* dataset (Giordani and Kiers, 2006) that in the following will help us to present the methodological proposal and its

properties. Data refer to a set of 16 different fruit juices that were submitted to a group of judges called to assign a score to the following six features: *Appearance*, *Smell*, *Taste*, *Naturalness*, *Sweetness*, *Density*.

The experts (judges) tried all juices and, for each juice, assigned scores to each one of the six features. Under this work hypothesis, the data matrix is defined by a three way data structure: *juices*×*features*×*judges*.

Many different analysis approaches exist to treat such kind of data: some approaches, such as the multiple factor analysis (Escofier and Pages, 1998) or the INDSCAL method (Carrol and Chang, 1970), are based on the treatment of the whole data matrix; alternative solutions consist of collapsing one dimension of the three matrix dimensions. For example all judges can be summarized into a mean or median score. The latter procedure leads to the more familiar *flat* $n \times p$ data matrix: *juices*×*features*.

To take into account the heterogeinety among judges single-valued data summarization can be replaced by the interval data coding: each interval score is, then, expression of both central tendency (midpoint) and intra-judges variability (range), and data are arranged in a $n \times 2p$ interval data flat matrix, where n refers to the number of observations and p to the number of interval-valued variables.

The generic $n \times p$ interval data matrix $[\mathbf{X}]$ has row $[\mathbf{x}]'_i$, with general term $[x]_{i,k} = [\underline{x}_{i,k}, \overline{x}_{i,k}]$, $i = 1, \ldots, n$ and $k = 1, \ldots, p$, with $\underline{x}_{i,k}$ and $\overline{x}_{i,k}$ the minimum and maximum observed values. The general term $[x]_{i,k}$ can be also represented as the *midpoint* $x^c_{i,k}$ and *range* (or *radius*) $x^r_{i,k}$ notation: $[x]_{i,k} = [\underline{x}_{i,k}, \overline{x}_{i,k}] = [x^c_{i,k} -$

Tab. 1: The *Juices* dataset

Juice	*Appearance*		*Smell*		*Taste*		*Naturalness*		*Sweetness*		*Density*	
Pineapple 1	6.61	7.66	5.82	6.66	6.18	7.31	5.45	6.85	5.63	6.75	3.92	4.98
Pineapple 2	6.75	7.59	5.90	7.30	5.65	6.98	5.23	6.56	5.52	6.92	3.28	4.69
Orange 1	6.75	7.59	7.12	8.24	6.39	7.44	5.67	6.72	5.83	6.67	3.64	4.97
Orange 2	6.89	7.45	6.06	6.90	6.89	7.94	5.60	6.72	6.01	7.13	3.88	4.93
Grapefruit 1	6.28	7.40	6.52	7.65	5.17	6.85	6.00	7.33	2.45	3.29	3.64	4.76
Grapefruit 2	6.31	7.43	5.63	6.75	6.35	7.47	6.11	7.23	4.14	5.19	3.06	4.46
Pear 1	6.92	7.76	7.19	8.24	7.14	8.19	6.44	7.49	7.70	8.54	7.22	8.27
Pear 2	7.62	8.18	6.32	7.44	7.73	8.57	6.79	7.63	7.78	8.62	6.83	7.67
Apricot 1	6.83	7.68	7.98	8.68	7.70	8.54	7.35	8.47	7.42	8.40	7.03	8.15
Apricot 2	7.32	8.16	7.21	8.19	5.17	6.71	4.66	6.06	4.90	6.31	5.79	6.77
Peach 1	7.09	7.93	6.94	7.78	6.42	7.54	5.70	7.10	6.69	7.68	5.20	5.90
Peach 2	6.98	7.82	6.22	7.06	7.54	8.38	6.88	7.72	6.83	7.81	5.01	5.85
Apple 1	6.82	7.52	5.47	6.59	7.42	8.40	5.66	7.20	7.37	8.29	5.90	6.74
Apple 2	6.60	7.72	6.28	7.40	6.31	7.43	5.72	7.12	6.81	7.65	5.47	6.59
Banana 1	4.96	6.37	3.92	5.60	3.64	5.32	4.27	5.95	4.76	6.16	3.62	4.74
Banana 2	5.27	6.67	3.68	5.36	3.26	4.94	3.92	5.46	4.23	5.91	3.65	4.77

$x^r_{i,k}, x^c_{i,k} + x^r_{i,k}]$. Midpoints and ranges are respectively defined by:

$$x^c = \tfrac{1}{2}(\overline{x} + \underline{x}), \qquad x^r = \tfrac{1}{2}(\overline{x} - \underline{x}).$$

In the midpoints/ranges notation, the matrix $[\mathbf{X}]$ is split in the matrices $\mathbf{X}^c$ and $\mathbf{X}^r$ that are called center and range matrix, respectively. It is worth to notice that midpoint-range and min-max interval data coding are equivalent. Table 1 shows interval data coding for the *Juices* dataset by the min-max notation. Hence, for example, the values 6.61 and 7.66 (in the first matrix cell) are the minimum and the maximum observed scores for the feature *Appearance* of the first pineapple juice.

From a geometric point of view, in dealing with single valued data, a p-variate statistical unit is represented by a dimensionless point for any p. Whereas in the interval p-dimensional Cartesian space, statistical units assume different geometric properties according to p. Each statistical unit is configurable as segment when $p = 1$, as a rectangle for $p = 2$, parallelepiped for $p = 3$ and, more generally, as a *parallelotope* when $p > 3$.

4. ARCHETYPES FOR INTERVAL DATA

In the framework of archetypal analysis, given the geometrical nature of interval data, in analogy with the single value case, the aim is to define some archetypal parallelotopes (that we will denote by $[\mathbf{A}]$), which should synthesize the locations and the shapes of all the other data. These archetypal parallelotopes are such that the others parallelotopes can be expressed as a convex combination of them, and they are a convex combination of all the others.

To define such new archetypes let us to consider each parallelotope be described by the midpoints and ranges coding. Hence, each statistical unit has coordinates into two linked multivariate spaces (the midpoint and range spaces). In such a case two sets of archetypes, $\mathbf{A}^c$ and $\mathbf{A}^r$, should be found in the midpoint and in the range spaces, respectively. As each parallelotope should be expressed as a unique convex combination of the archetypal parallelotopes in terms of midpoints and ranges, an additional constraint is imposed: the mixture coefficients α'_i in eqn. (1) should be the same in the two spaces. The α'_i coefficients represent the algebraic linkage of the two optimizations, and hence the linkage between the two spaces. In order to define the parallelotopes-archetypes the least square criterion in eqn. (5) and (6) has to be rewritten in terms of intervals. In such a case the euclidean metric has to be replaced by an appropriate one to treat intervals. In

this paper we propose to use the Hausdorff metric, which was proposed by Felix Hausdorff in the early of 20^{th} century as a measure of distance between compact subsets in $\mathbb{R}^p$.

Given two closed sets $S \subset \mathbb{R}^p$ and $T \subset \mathbb{R}^p$ and a metric $d(\cdot)$, the distance *from* a point $x \in \mathbb{R}^p$ *to* the subset S is defined as:

$$d(x,S) = \min_{\tilde{s} \in S} d(x,\tilde{s}).$$

Let us define the quantities $h(S,T)$ and $h(T,S)$ as:

$$\begin{aligned} h(S,T) &= \max_{\tilde{s} \in S} d(\tilde{s},T) = \max_{\tilde{s} \in S} \min_{\tilde{t} \in T} d(\tilde{s},\tilde{t}) \\ h(T,S) &= \max_{\tilde{t} \in T} d(\tilde{t},S) = \max_{\tilde{t} \in T} \min_{\tilde{s} \in S} d(\tilde{t},\tilde{s}). \end{aligned} \tag{7}$$

The Hausdorff distance $H(S,T)$ between S and T is defined as:

$$H(S,T) = \max\left(h(S,T), h(T,S)\right). \tag{8}$$

In the special case of $\mathbb{R}$, the compact sets are intervals and, hence, $S = [\underline{s}, \overline{s}]$ and $T = [\underline{t}, \overline{t}]$. The Hausdorff distance between these two generic intervals is given by:

$$H(S,T) = \max\{|\,\overline{s} - \overline{t}\,|, |\,\underline{s} - \underline{t}\,|\}, \tag{9}$$

and it is easy to show that the Hausdorff distance can be written in terms of centers and ranges as follows:

$$H(S,T) = |\,s^c - t^c\,| + |\,s^r - t^r\,|\,. \tag{10}$$

Furthermore, we have that $H(S,T) \geq 0$ and $H(S,T) = H(T,S)$. In addition, let U be a generic compact subset in $\mathbb{R}$, the triangular inequality $H(S,U) \leq H(S,T) + H(T,U)$ can be easily proved taking into account the definition of distance in (8) (Neumaier, 1990; Palumbo and Irpino, 2005).

The generalization of the Hausdorff distance in $\mathbb{IR}^p$ is very complex. However, when the compact subsets are restricted to some special cases, as the parallelotopes we are considering in this paper, the Hausdorff metric can be easily generalized. In such a case it can be proved that the distance between two parallelotopes in $\mathbb{R}^p$ is the sum of the Hausdorff distances in each dimension.

Reverting to our aim, given the interval data matrix $[\mathbf{X}]$ and two parallelotopes $[\mathbf{x}]'_i$ and $[\mathbf{x}]'_{i'}$ in $\mathbb{IR}^p$, the Hausdorff distance between them is:

$$\begin{aligned} H([\mathbf{x}]'_i, [\mathbf{x}]'_{i'}) &= \sum_{k=1}^{p} \max\left\{|\overline{x}_{ik} - \overline{x}_{i'k}|, |\underline{x}_{ik} - \underline{x}_{i'k}|\right\} \\ &= \sum_{k=1}^{p} \left(|x^c_{ik} - x^c_{i'k}| + |x^r_{ik} - x^r_{i'k}|\right). \end{aligned} \tag{11}$$

Given m, with m the number of archetypes to be derived, and the Hausdorff distance defined above, in analogy with eqn. (6), the parallelotope-archetypes $[\mathbf{A}]$ are such that they minimize the quantity $HRS(m)$, the sum over i of the Hausdorff distances among the parallelotopes $[\mathbf{x}]'_i$ and their representation in terms of the archetypes $[\tilde{\mathbf{x}}]'_i$:

$$HRS(m) = \sum_{i=1}^{n} H\left([\mathbf{x}]'_i, [\tilde{\mathbf{x}}]'_i\right) \tag{12}$$

where $[\tilde{\mathbf{x}}]'_i = \alpha'_i[\mathbf{A}]$ is the best approximation of $[\mathbf{x}]'_i$ through a convex combination of the archetypes, and where α_i indicates the weight of the archetypes in determinig the i-th parallelotope. Considering the midpoints and ranges notation, eqn.(12) can be rewritten as follows:

$$HRS(m) = \sum_{i=1}^{n}\sum_{k=1}^{p}\left(\left|x^c_{ik} - \sum_{j=1}^{m}\alpha_{ij}a^c_{jk}\right| + \left|x^r_{ik} - \sum_{j=1}^{m}\alpha_{ij}a^r_{jk}\right|\right), \tag{13}$$

where α_{ij} indicates the weight of j-th archetype on the i-th statistical unit, a^c_{jk} and a^r_{jk} are the general terms of the matrices $\mathbf{A}^c$ and $\mathbf{A}^r$ indicating the k-th coordinates of the j-th archetype in the midpoints and ranges spaces.

In analogy with the single value case, the quantity $HRS(m)$ has to be minimized under the constraints:

i) $\alpha_{ij} \geq 0 \quad \forall i, j \qquad \alpha'_i\mathbf{1} = 1 \quad \forall i,$

ii) $\mathbf{A}^c = \mathbf{B}^c\mathbf{X}^c$ and $\mathbf{A}^r = \mathbf{B}^r\mathbf{X}^r$,

iii) $0 \leq \beta^c_{ji} \leq 1$ and $0 \leq \beta^r_{ji} \leq 1$, where β^c_{ji} and β^r_{ji} are the general terms of matrices $\mathbf{B}^c$ and $\mathbf{B}^r$,

iv) $\mathbf{B}^c\mathbf{1}_n = \mathbf{1}_m$ and $\mathbf{B}^r\mathbf{1}_n = \mathbf{1}_m$, with $\mathbf{1}_n$ and $\mathbf{1}_m$ the unitary vector of order n and m.

where *ii*), *iii*) and *iv*) impose archetypes in the space of midpoints and ranges must be defined as convex combinations of the original data.

It is worth noticing that the interval data archetypal problem consists in defining two archetype matrices $\mathbf{A}^c$ and $\mathbf{A}^r$ and one system of weights α'_i, $i = 1, \ldots, n$, that is common to midpoints and ranges. Moreover, the Hausdorff distance implies the use of the L_1 norm. Consequently, the alternate least squares algorithm that has been used in the original archetypal analysis would not be able to determine the solution. Hence, in our proposal the minimization problem to determine

the parallelotopes-archetypes can be solved via the mathematical programming approach.

4.1 AN ILLUSTRATIVE APPLICATION TO THE JUICES DATASET

This section presents some analytical and graphical results obtained starting from the *Juices* dataset. Looking at behavior of the $HRS(m)$ function we selected three parallelotope-archetypes, denoted by $[\mathbf{a}]_1, [\mathbf{a}]_2$ and $[\mathbf{a}]_3$.

Table 2 reports the α_{ij} coefficients. The archetype $[\mathbf{a}]_1$ corresponds to the second banana juice, and it is also very close to the other banana juice. The archetype $[\mathbf{a}]_2$ is the first orange juice, while the third archetype corresponds to an apricot and to a pear juice. The α_{ij}'s can be used to obtain a fuzzy clustering of the juices based on preferences, assigning each juice to the archetypes for which the α_{ij} is maximum.

Tab. 2: α_{ij}'s coefficients for the *Juice* dataset for *m* = 3 archetypes. a1 a2 a3 Sum

	α_1	α_2	α_3	Sum
Pineapple 1	0,414	0,373	0,213	1
Pineapple 2	0,34	0,51	0,15	1
Orange 1	0,01	**0,99**	0,00	1
Orange 2	0,32	0,32	0,36	1
Grapefruit 1	0,35	0,65	0,00	1
Grapefruit 2	0,45	0,48	0,07	1
Pear 1	0,02	0,20	0,78	1
Pear 2	0,00	0,00	**1,00**	1
Apricot 1	0,00	0,00	**1,00**	1
Apricot 2	0,38	0,62	0,00	1
Peach 1	0,05	0,67	0,28	1
Peach 2	0,174	0,162	0,664	1
Apple 1	0,30	0,00	0,70	1
Apple 2	0,195	0,348	0,457	1
Banana 1	0,995	0,003	0,002	1
Banana 2	**1,00**	0,00	0,00	1

In analogy to the proposal by Porzio *et al*. (2006), the archetypes can be represented through some appropriate iconic plots. In such a case we propose to visualize them by the stars (Hartigan, 1975), in their version for interval data (Noirhomme-Fraiture, 2002). In Figure 1 the three archetypes for the *Juices* dataset are represented. The first archetype $[\mathbf{a}]_1$ (the inner blu star-band) repre-

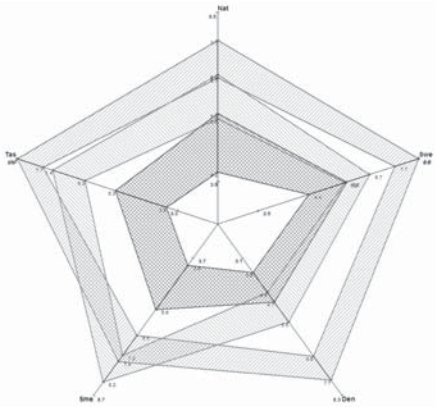

Fig. 1: Three interval archetypes for the *Juices* dataset represented through the star for interval data.

sents juices with low and spread preferences on all the features, while the third archetype $[\mathbf{a}]_3$ summarizes juices with highest preferences (the outer green star). Finally, the second archetytpe $[\mathbf{a}]_2$ (the purple star) stands for juice with intermediate preferences, except for the smell attribute that has the highest scores.

The relative positions and distances of data with respect to the archetypes in terms of the α_{ij}'s are displayed in Figure 2. The linear constraint imposed on the α_{ij}'s weights allows us to plot them on a bivariate plan instead of a three dimensional space. Coordinates are determined by subtracting the third column to the first two ones. The juices at the center of the triangle cannot be easily classified with respect to the archetypes.

The analysis of Table 3, which contains the $\mathbf{B}^c$ and $\mathbf{B}^r$ coefficient matrices for the *Juices* dataset for $m = 3$ archetypes, highlights the different role played by the observed data in determining the archetypes in the two spaces. We note that β^c_{ij} and β^r_{ij} coefficients can be interpreted similarly to the absolute contributions in the principal component analysis. For example, in the center space the first orange juice heavily contributes to $[\mathbf{a}]_2$, while the second orange juice does not contribute at all; at the same time both juices contribute the same archetype in a similar way in the range space. This means that the $[\mathbf{a}]_2$ shape will be a mix of the two orange juice shapes. A similar remarks could be done for the banana juices and the third archetype: in such a case the $[\mathbf{a}]_3$ location will closely resemble the *Banana 1* location, while its shape will be a mix of the two juice shapes. It is worth to note that there are some data that do not contribute at all to any archetypes.

In Table 4 there are the coordinates of the three interval-valued archetypes,

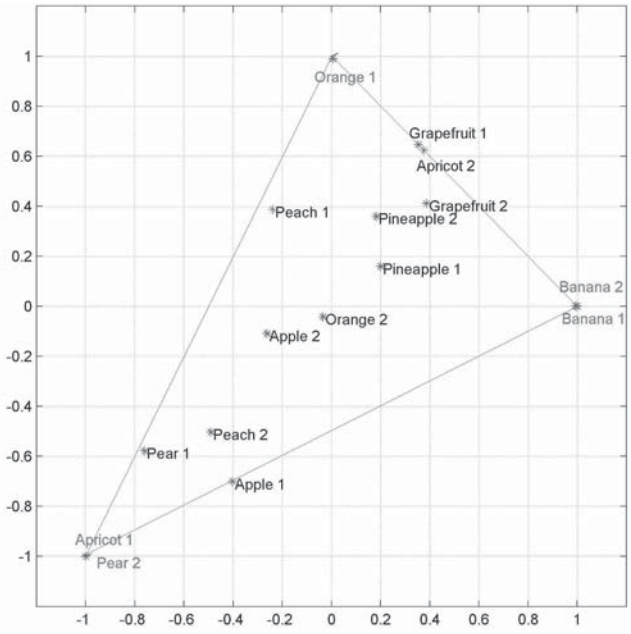

Fig. 2: Plot of the observed data using as coordinates the α'_i coefficients for the *Juices* dataset for $m = 3$ archetypes.

computed through a minimization procedure taking into account midpoints and ranges.

A graphical analysis of archetypes can be performed through parallel coordinates plot (Inselberg, 1985; Wegman, 1990; Wegman and Qiang, 1997) by jointly visualizing the archetypes and original data values (for a discussion on the use of parallel coordinates to graphical exploration of archetypes see Porzio *et al*., 2006). Displaying data in such a way, each statistical unit is compared to the archetypes and is evaluated in terms of its deviations from the archetypes.

However, interval data would make much less effective the parallel coordinates display. This proposal presents a joint display of original data and archetypes on the first *Midpoint-Range* Principal Component Analysis (MR-PCA) (Palumbo and Lauro, 2003; Lauro and Palumbo, 2005). Details of the MR-PCA pass the scope of the present paper and we refer to the quoted papers for them. It is enough to say the interpretation can be done like in the classical PCA. Dealing with interval variables, statistical units are compared in terms of position, as in the standard case, and in terms of *size* and *shape*.

The total inertia associated to the first factorial plan in Figure 3 is equal to

Tab. 3 – B^c **and** B^r **coefficient matrices for the *Juices* dataset for *m*=3 archetypes.**

	Center Space			*Range Space*		
	β_1^c	β_2^c	β_3^c	β_1^r	β_2^r	β_3^r
Pineapple 1	0.000	0.000	0.000	0.000	0.095	0.000
Pineapple 2	0.000	0.000	0.000	0.000	0.000	0.000
Orange 1	0.000	0.727	0.000	0.000	0.225	0.000
Orange 2	0.024	0.000	0.000	0.000	0.159	0.000
Grapefruit 1	0.000	0.013	0.000	0.118	0.084	0.000
Grapefruit 2	0.086	0.000	0.000	0.000	0.025	0.000
Pear 1	0.000	0.000	0.031	0.000	0.082	0.000
Pear 2	0.000	0.022	0.830	0.000	0.017	0.347
Apricot 1	0.000	0.070	0.079	0.000	0.161	0.226
Apricot 2	0.000	0.168	0.000	0.000	0.000	0.000
Peach 1	0.000	0.000	0.000	0.000	0.004	0.000
Peach 2	0.000	0.000	0.000	0.000	0.000	0.287
Apple 1	0.000	0.000	0.060	0.000	0.085	0.123
Apple 2	0.000	0.000	0.000	0.000	0.063	0.017
Banana 1	0.148	0.000	0.000	0.480	0.000	0.000
Banana 2	0.742	0.000	0.000	0.402	0.000	0.000
	1	**1**	**1**	**1**	**1**	**1**

73.75%. Parallelotope archetypes have been projected as supplementary statistical units.

It is extremely important to remark the consistency between the results of the previuos analysis and the representation obtained through the MR-PCA. Notice that archetypes are very close to those units that have the highest α_i' coefficients. Looking at the archetype $[\mathbf{a}]_3$, on the left side of the display, it is worth noticing how it is very close and pretty much similar to *Pear 2* and *Apricot 1*. Moreover, even if *Pear 2* is closer to $[\mathbf{a}]_3$ than *Apricot 1*, shape of $[\mathbf{a}]_3$ resembles more closely the *Apricot 1* shape.

On the opposite side, the two banana juices along with $[\mathbf{a}]_1$ clusterize tightly

Tab. 4 – Interval archetypes for the *Juices* dataset for ***m* = 3** archetypes.

Juice	*Appearance*		*Smell*		*Taste*		*Naturalness*		*Sweetness*		*Density*	
$[\mathbf{a}]_1'$	5,34	6,72	3,94	5,6	3,64	5,3	4,2	5,7	4,4	5,85	3,61	4,73
$[\mathbf{a}]_2'$	6,86	7,71	7,22	8,2	6,31	7,4	5,6	6,8	5,8	6,73	4,37	5,49
$[\mathbf{a}]_3'$	7,42	8,15	6,5	7,4	7,69	8,5	6,7	7,7	7,7	8,61	6,78	7,69

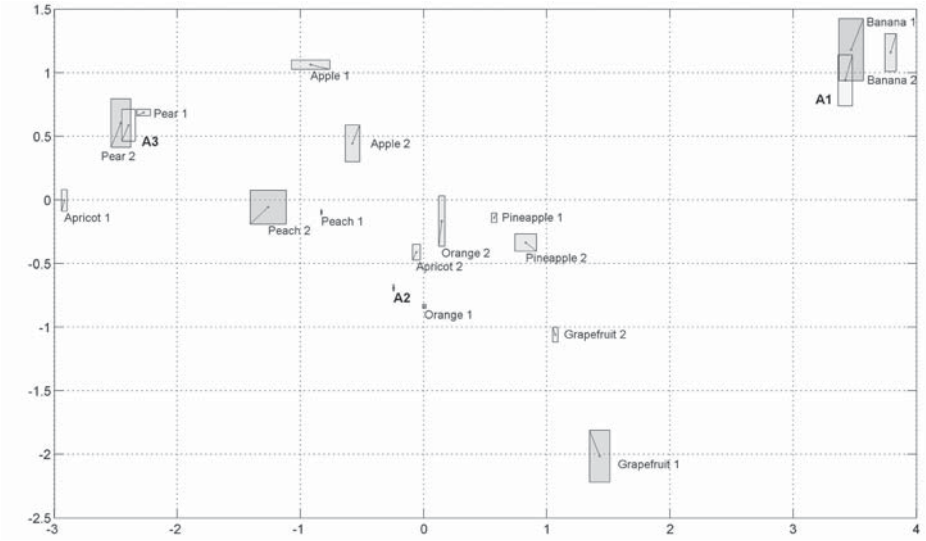

Fig. 3: Archetypes and original statistical units over the first PCA factorial plan.

together, for both location and shape. While the $[\mathbf{a}]_2$ archetype, as it falls around the origin of the axes, is not well represented, and hence its position and shape cannot be interpreted.

By our results the juices seem to clusterize for types of fruits and preference levels. However, to improve the interpretation additional information, such as the brand, the price and so on, should be considered.

5. SOME CONCLUDING REMARKS

The results tend to confirm that the proposed method, that extends archetypal analysis from single value data to interval data, can be promisingly applied in marketing research.

At same time we believe that there is the possibility to use the proposed method in other applicative fields. More testing on real data is therefore necessary.

REFERENCES

ALLENBY, G.M., GINTER, J.L. (1995), Using Extremes to Design Products and Segment Markets, *Journal of Markenting Research*, **32**, 392-403.

ANDERSON, L., WEINER, J.L. (2004), *Actionable Market Segmentation Guaranteed (Part Two)*. Knowledge Center Ipsos-Insight (*www.ipsos.com*).

CARROLL,J.D., CHANG, J.J. (1970), Analysis of individual differences in multidimensional scaling via an N-way generalisation of the "Eckart-Young" decomposition, *Psychometrika*, **35**, 282-319.

CHAN, B.H.P.,MITCHELL, D.A. and CRAML.E. (2003),Archetypal analysis of galaxy spectra. *Monthly Notice of the Royal Astronomical Society*, **338**, 3, 790–795.

CUTLER, A., BREIMAN, L. (1994), Archetypal Analysis.n*Technometrics*, **36**, 338–347.

DE SARBO, W., WAGNER, A.K., WEDEL, M. (2004), Applications of Multivariate Latent Variable Models in Marketing, in Wind J.(Ed.) *Advances in Marketing Research andModeling: The Academic and Industry Impact of Paul E. Green*, Boston, MA, Kluwer, 43-67.

D'ESPOSITO, M.R., RAGOZINI, G. (2004), Multivariate Ordering in Performance Analysis. In: *Atti XLII Riunione Scientifica SIS*. CLEUP, Padova, 51–55.

ELDER, A., PINNEL,J. (2003), *Archetypal Analysis: an Alternative Approach to Finding and Defining Segments, 2003 Sawtooth Software Conference Proceedings*, Sequim, WA, 113-129.

ESCOFIER, B., PAGÉS, J. (1998), *Analyses factorielles multiples*, Dunod, Paris.

GIORDANI, P., KIERS, H.A.L. (2006), A comparison of three methods for principal component analysis of fuzzy interval data, *Computational Statistics and Data Analysis*, **51**, 379-397.

HARTIGAN, J.A. (1975), Printer Graphics for Clustering, *Journal of Statistical Computation and Simulation*, **4**, 187–213.

INSELBERG, A. (1985), The Plane With Parallel Coordinates. *The Visual Computer*, **1**, 69–91.

LAURO, C. N., PALUMBO, F. (2005), Principal Component Analysis for Non-Precise Data, in Vichi M. et al. Eds. 'New Developments in Classification and Data Analysis', *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Heidelberg, 173–184.

LI, S.,WANG, P., LOUVIERE, J., CARSON, R. (2003), ArchetypalAnalysis: 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*, Adelaide,1674-1679.

MORRIS, L., SCHMOLZE, R. (2006), Consumer Archetypes: a New Approach to Developing Consumer Understanding Frameworks, *Journal of Marketing Research*, **46**, 289-300.

NEUMAIER, A. (1990), *Interval Methods for systems of Equations*, Cambridge University Press, Cambridge.

NOIRHOMME-FRAITURE, M. (2002), Visualization of Large Data Sets: The Zoom Star Solution, *The Electronic Journal of Symbolic Data Analysis*, **0**, 1–14.

PALUMBO, F., IRPINO, A. (2005), Multidimensional Interval-Data: Metrics and Factorial Analysis, in Jacques Janssen and Philippe Lenca Eds., *Proceeding of ASMDA' 05 conference*, Brest, May 2005, 689-698.

PALUMBO, F., LAURO, C. N. (2003),A PCA for interval valued data based onmidpoints and radii, *in*'H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J.Meulman, eds, 'New developments in Psychometrics', Psychometric Society, Springer-Verlag, Tokyo.

PORZIO, G.C., RAGOZINI, G., VISTOCCO, D. (2006), Archertypal Analysis for Data Driven Benchmarking , in Zani *et al.* Eds., *Data Analysis, Classification and the Forward Search*, Spinger-Verlag, Heidelberg, 309-318.

RATNER, B. (2003), *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, Chapman and Hall/CRC, New York.

RIEDESEL, P. (2003), Archetypal Analysis in Marketing Research: A New Way of Understanding Consumer Heterogeneity, http://www.action-research.com/archtype.html.

STONE, E. (2002), Exploring Archetypal Dynamics of Pattern Formation in Cellular Flames. *Physica D* , **161**, 163–186.

STONE, E., CUTLER, A. (1996), Introduction to Archetypal Analysis of Spatio-temporal Dynamics. *Physica D*, **96**, 110–131.

WEGMAN, E.J. (1990), Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, **85**, 664–675.

WEGMAN, E.J., QIANG, L. (1997), High dimensional clustering using parallel coordinates and the grand tour, *Computing Science and Statistics*, **28**, 352–360.

ANALISI DEGLI ARCHETIPI PER DATI INTERVALLARI NELLE RICERCHE DI MERCATO

Riassunto

Usualmente nelle ricerche di mercato, un obiettivo è l'individuazione di gruppi e segmenti di prodotti e/o consumatori. Tuttavia i metodi classici possono produrre risultati poco interpretabili, poichè tendono ad individuare consumatori o prodotti medi, che spesso non sono molto diversificati fra loro. Nelle ricerche di mercato, quindi, con l'obiettivo di trovare segmenti ben separati ed estremi, è stata introdotta l'analisi degli archetipi. D'altro canto, notiamo che spesso le caratteriche dei prodotti e le preferenze dei consumatori potrebbero essere espresse più adeguatamente attraverso intervalli di valori. Per coniugare queste due esigenze, in questo articolo, proponiamo una estensione della analisi degli archetipi per dati di tipo intervallare, fornendone una definizione analitica in termini di distanza di Hausdorff, analizzandone le caratteristiche geometriche ed indicando alcuni strumenti di visualizzazione per dati ad intervallo. Nell'articolo viene presentata anche un'applicazione a dati di preferenza.