Interval Archetypes: A New Tool for Interval Data Analysis

Article *in* Statistical Analysis and Data Mining · August 2012

DOI: 10.1002/sam.11140

3 authors:

maria rosaria d'esposito
Università degli Studi di Salerno
11 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Francesco Palumbo
University of Naples Federico II
46 PUBLICATIONS **192** CITATIONS
SEE PROFILE

Giancarlo Ragozini
University of Naples Federico II
34 PUBLICATIONS **62** CITATIONS
SEE PROFILE

Interval Archetypes: A New Tool for Interval Data Analysis

Maria R. D'Esposito[1], Francesco Palumbo[2*] and Giancarlo Ragozini[3]

[1]*Department of Economics and Statistics, University of Salerno, Fisciano, Italy*

[2]*Department of Theory and Methods of the Human and Social Sciences, Federico II University of Naples, Naples, Italy*

[3]*Department of Sociology "Gino Germani", Federico II University of Naples, Naples, Italy*

Received 16 September 2011; revised 14 December 2011; accepted 4 February 2012
DOI:10.1002/sam.11140
Published online 22 March 2012 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Archetypal analysis aims at synthesizing single-valued data sets through a few (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. In this paper, we extend this methodology to the case of interval-valued data, which represent a special case of set-valued data, where the sets are compact and identified by ordered pairs of values. In addition, we propose to use interval archetypes as a tool in an analysis strategy to explore and mine complex data sets. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 5: 322–335, 2012

Keywords: archetypal analysis; interval-valued data; barycentric coordinates; exploratory data analysis; graphical data analysis

1. INTRODUCTION

In standard data analysis, statistical units are described by single-valued variables which are usually represented as points in a multidimensional space. Symbolic data constitute a wider concept of data, where the variables can assume different types of values which can be weighted and linked by logical rules.

According to this definition, interval-valued data are a special kind of symbolic data. They can result from measurement errors, or as the output of queries to databases—where two extreme values (minimum and maximum, as well as a pair of symmetric quantiles with respect to the median value) are queried instead of the average value—or they can be derived to synthesize large sets of observations [1]. Moreover, a very interesting condition arises when data are naturally interval-valued. This is the typical situation when describing species or giving a product specification.

Correspondence to: F. Palumbo (fpalumbo@unina.it)

This paper focuses on archetypal analysis for interval data. In the classical data setup, archetypal analysis [2] aims at synthesizing a set of n statistical units, described by p variables, through a set of (not necessarily observed) points that are called archetypes. The latter are derived under the constraints that all points are represented as their convex combination, and that they are a convex combination of the data. The archetypes are located on the boundary of the data convex hull, and can, therefore, provide an outward–inward point of view on the data scatter. This will allow the analyst to explore the data cloud peripheries and highlight many data patterns such as small groups, gaps in the data structure, outlying values, asymmetries and irregularities in the data shape. Archetypes can, hence, be a powerful tool for visual exploratory data analysis. Up to now, archetypal analysis has been applied in many fields. In the field of physics, it has been used to detect clusters of cellular flames [3,4] and of galaxy spectra [5]. It has found application as a tool for image decomposition [6,7], where archetypal analysis seems to provide results which are easily interpretable in terms of physical meaning. Marketing research is also a relevant field of application. Archetypes

have been associated with the idea of archetypal consumers [8] and, to overcome the drawbacks of the classical segmentation technique, they have been exploited for market segmentation and consumer fuzzy clustering [9,10]. In particular, they have been used to obtain extreme and well-separated consumer profiles. In performance analysis, archetypes have been exploited to construct data-driven benchmarks [11], to analyze the performance of central processing unit (CPU) [12] and to obtain a multivariate ordering procedure based on the idea of the 'worst-best' direction selected through the archetypes [13].

Given all these features and applications, it is our opinion that extending the methodology of archetypal analysis to the case of interval-coded data would result in a powerful tool for the latter type of data. Our interest in this type of analysis is twofold. On the one hand, we study archetypes in the interval data setting for their capabilities to extract and analyze valuable information; on the other hand, according to the two-level paradigm [14], we analyze how to use interval archetypes to define higher order symbolic objects. This paper focuses on the first issue; readers interested in archetypal analysis interpretation in the framework of symbolic objects are referred to ref. 15.

The paper is organized as follows: Sections 2 and 3 present a brief review of archetypal analysis for single-valued data and interval data definitions, respectively. In Section 4, our proposal of archetypes for interval-coded data is introduced. A discussion of the use of interval archetypes to explore and to analyze interval data is provided in Section 5, and Section 6 presents an application to a real data set along with a discussion of some computational issues.

2. AN OVERVIEW OF ARCHETYPAL ANALYSIS FOR SINGLE-VALUED DATA

Archetypal analysis relies on the idea of 'pure individual types' (the archetypes), a few points lying on the boundary of the data scatter and characterizing the archetypal pattern in the data.

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

$$\boldsymbol{\beta}_j(m) = (\beta_{j1}(m), \ldots, \beta_{jn}(m))' \quad \text{and}$$
$$\boldsymbol{\gamma}_i(m) = (\gamma_{i1}(m), \ldots, \gamma_{im}(m))',$$

the archetypes $\mathbf{a}_j(m)$, $j = 1, \ldots, m$, are defined as the p-vectors that satisfy the following conditions:

$$\mathbf{a}'_j(m) = \boldsymbol{\beta}'_j(m)\mathbf{X}, \quad j = 1, \ldots, m, \quad \beta_{ji}(m) \geq 0 \,\forall j, i,$$
$$\boldsymbol{\beta}'_j(m)\mathbf{1} = 1 \quad \forall j; \tag{1}$$
$$\mathbf{x}'_i = \boldsymbol{\gamma}'_i(m)\mathbf{A}(m), \quad i = 1, \ldots, n, \quad \gamma_{ij}(m) \geq 0 \,\forall i, j,$$
$$\boldsymbol{\gamma}'_i(m)\mathbf{1} = 1 \,\forall i. \tag{2}$$

An exact solution of Eqs. (1) and (2) exists only if $m = V$, where V indicates the cardinality of the set of the vertices of the data convex hull, and if the archetypes coincide with the V vertices of the data convex hull [11].

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

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

Define

$$\tilde{\mathbf{X}}(m) = \left(\tilde{\mathbf{x}}_1(m), \ldots, \tilde{\mathbf{x}}_n(m)\right)', \qquad \tilde{\mathbf{X}}(m) \in \Re^{n\times p}, \tag{3}$$
$$\boldsymbol{\Gamma}(m) = \left(\boldsymbol{\gamma}_1(m), \ldots, \boldsymbol{\gamma}_n(m)\right)', \qquad \boldsymbol{\Gamma}(m) \in \Re^{n\times m}, \tag{4}$$
$$\mathbf{B}(m) = \left(\boldsymbol{\beta}_1(m), \ldots, \boldsymbol{\beta}_m(m)\right), \qquad \mathbf{B}(m) \in \Re^{n\times m}; \tag{5}$$

and

$$\begin{aligned} \mathrm{RSS}(m) &= \|\mathbf{X} - \tilde{\mathbf{X}}(m)\|_F \\ &= \|\mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{A}(m)\|_F \\ &= \left\|\mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{B}'(m)\mathbf{X}\right\|_F, \end{aligned} \tag{6}$$

where $\mathrm{RSS}(m)$ is the residual sum of squares given m archetypes, and $\|\mathbf{Y}\|_F = \sqrt{Tr\,(\mathbf{YY}')}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$.

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

$$\min_{\boldsymbol{\Gamma}(m),\mathbf{B}(m)} \mathrm{RSS}(m) = \min_{\boldsymbol{\Gamma}(m),\mathbf{B}(m)} \left\|\mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{B}'(m)\mathbf{X}\right\|_F \tag{7}$$

holding all the other conditions on the coefficients $\boldsymbol{\beta}'_j(m)$ and $\boldsymbol{\gamma}'_i(m)$[1].

[1] Originally, Cutler and Breiman [2] defined the archetypes as those points, a mixture of the observed data, that minimize the residual sum of square function for a fixed m given constraints on the mixture coefficients in the euclidean metric. Our presentation is different even if it leads to the same results: it is thought to emphasize the geometric features of the archetypal analysis useful for our purposes.

The solution to this minimization problem depends on m, and solutions are not nested as m varies. That is, the archetypal points that solve the minimization problem in Eq. (7) for $m = m^*$ are not necessarily a subset of the solutions for $m = m^* + 1$. By denoting with $\mathbf{a}'_j(m)$ the j-th archetype for a given m, generally $\mathbf{a}'_j(m) \neq \mathbf{a}'_j(l)$, for $m \neq l$. For each m, as m increases, the archetypes change to capture the shape of the convex hull of data better.

The RSS(m) in Eq. (6) is a decreasing function of m that has the maximum for $m = 1$ and goes to zero for m approaching the number of vertices of the convex hull. For a given m, RSS(m) in Eq. (6) highlights the synthesizing power of the archetypes, as it shows how well their convex combination approximates the given data points.

3. INTERVAL DATA: NOTATIONS AND DEFINITIONS

Interval data represent a special case of set-valued data, where the sets are compact and identified by ordered pairs of values:

$$\mathbb{s} = [\underline{s}, \overline{s}] = \left\{ s : s \in \Re, \underline{s} \leq s \leq \overline{s} \right\},$$

where $\underline{s}$ and $\overline{s}$ are the interval bound values. According to the most widely used notation in interval data analysis, the set of all intervals is usually written as $\Im\Re$ [16].

Given an interval $\mathbb{s} \in \Im\Re$, and the quantities *midpoint* and *range* defined respectively as:

$$\check{s} = \frac{1}{2}(\underline{s} + \overline{s}),$$
$$\Delta s = \frac{1}{2}(\overline{s} - \underline{s}),$$

the interval is equivalently defined as

$$\mathbb{s} = [\underline{s}, \overline{s}] = [\check{s} - \Delta s, \check{s} + \Delta s].$$

In p dimensions, an interval is defined through the Cartesian product of p intervals $\mathbb{s} \in \Im\Re$. A p-dimensional interval is represented as a parallelotope, and the set of intervals in p dimensions is written $\Im\Re^p$ [16].

Let $\mathbb{S}$ be an $n \times p$ matrix whose elements $\mathbb{s}_{ij}$ are intervals in $\Im\Re$:

$$\mathbb{S} = \begin{bmatrix} \mathbb{s}_{11} & \cdots & \mathbb{s}_{1p} \\ \vdots & \mathbb{s}_{ij} & \vdots \\ \mathbb{s}_{n1} & \cdots & \mathbb{s}_{np} \end{bmatrix} = (\mathbb{s}_{ij}),$$

where each $\mathbb{s}_{ij} = [\underline{s}_{ij}, \overline{s}_{ij}]$. The matrix $\mathbb{S}$ is an interval matrix, and it can be written also as

$$\mathbb{S} = [\underline{\mathbf{S}}, \overline{\mathbf{S}}], \text{ with } \underline{\mathbf{S}} = (\underline{s}_{ij}) \text{ and } \overline{\mathbf{S}} = (\overline{s}_{ij}), \quad \underline{\mathbf{S}} \leq \overline{\mathbf{S}}.$$

The set of $n \times p$ interval matrices is written as $\Im\Re^{n \times p}$ [16].

It is also possible to write the interval matrix through the midpoint-range notation as

$$\mathbb{S} = \left[\check{\mathbf{S}} - \Delta\mathbf{S}, \check{\mathbf{S}} + \Delta\mathbf{S} \right], \tag{8}$$

where $\check{\mathbf{S}}$ is the matrix of midpoints of the interval elements in $\mathbb{S}$, and $\Delta\mathbf{S}$ is the matrix of ranges of the interval elements in $\mathbb{S}$. They are defined as

$$\check{\mathbf{S}} = \frac{1}{2}(\underline{\mathbf{S}} + \overline{\mathbf{S}}), \quad \Delta\mathbf{S} = \frac{1}{2}(\underline{\mathbf{S}} - \overline{\mathbf{S}}). \tag{9}$$

Operations between interval matrices are formally defined as the corresponding ones between single-valued matrices. The product between an interval matrix and a single-valued matrix can be also defined [16].

4. ARCHETYPAL ANALYSIS FOR INTERVAL-CODED DATA

Let $\mathbb{X} \in \Im\Re^{n \times p}$ be an interval matrix, $\mathbb{X} = (\mathbb{x}_1, \ldots, \mathbb{x}_n)'$, $\mathbb{X} = \left[\underline{\mathbf{X}}, \overline{\mathbf{X}}\right] = \left[\check{\mathbf{X}} - \Delta\mathbf{X}, \check{\mathbf{X}} + \Delta\mathbf{X}\right]$. Each observed data point $\mathbb{x}_i$ is now a parallelotope. In analogy with the single-valued case presented in Section 2, we state that the aim of interval archetypal analysis should be to find m archetypal parallelotopes, denoted by $\mathbb{a}_1, \ldots, \mathbb{a}_m$, with $(\mathbb{a}_1, \ldots, \mathbb{a}_m)' = \mathbb{A}(m)$, $\mathbb{A}(m) \in \Im\Re^{m \times p}$, such that:

$$\mathbb{A}(m) = \mathbf{B}'(m)\mathbb{X}, \tag{10}$$
$$\mathbb{X} = \mathbf{\Gamma}(m)\mathbb{A}(m), \tag{11}$$

where the matrices $\mathbf{B}$ and $\mathbf{\Gamma}$ are defined as in Eqs (4) and (5). The interval archetypes are such that the observed parallelotopes $\mathbb{x}_i$'s can be expressed as a convex combination of them, and they are a convex combination of all the data points $\mathbb{x}_i$'s. These archetypal parallelotopes synthesize the locations and the shapes of all the other data.

To find these new interval archetypes, let us consider the midpoint and range notation (Eqs. (8) and (9)). Finding the matrix $\mathbb{A}(m)$ of the archetypal parallelotopes is equivalent to finding two archetypal matrices $\check{\mathbf{A}}(m)$ and $\Delta\mathbf{A}(m)$, $\check{\mathbf{A}}(m) = (\check{\mathbf{a}}_1, \ldots, \check{\mathbf{a}}_m)'$, $\Delta\mathbf{A}(m) = (\Delta\mathbf{a}_1, \ldots, \Delta\mathbf{a}_m)'$, with

$$\mathbb{A}(m) = [\check{\mathbf{A}}(m) - \Delta\mathbf{A}(m), \check{\mathbf{A}}(m) + \Delta\mathbf{A}(m)],$$
$$\check{\mathbf{A}}(m) \in \Re^{m \times p}, \quad \Delta\mathbf{A}(m) \in \Re^{m \times p},$$

such that

$$\check{\mathbf{A}}(m) = \mathbf{B}'(m)\check{\mathbf{X}}, \tag{12a}$$

$$\check{\mathbf{X}} = \mathbf{\Gamma}(m)\check{\mathbf{A}}(m), \tag{12b}$$

and

$$\Delta\mathbf{A}(m) = \mathbf{B}'(m)\Delta\mathbf{X}, \tag{13a}$$

$$\Delta\mathbf{X} = \mathbf{\Gamma}(m)\Delta\mathbf{A}(m), \tag{13b}$$

in the midpoint and in the range spaces, respectively. As each observed parallelotope $\mathbb{x}_i$ should be expressed as a convex combination of the archetypes in terms of midpoints and ranges, we impose the additional constraint that the convex combination coefficient matrices $\mathbf{B}(m)$ and $\mathbf{\Gamma}(m)$, defined in Eqs. (4) and (5), should be the same in the two spaces. These coefficients represent the algebraic linkage between the two spaces.

To find the parallelotope archetypes, in analogy with the classical case, we relax Eq. (11) or, equivalently, Eq. (12b) and (13b), and we rewrite the least square criterion in Eq. (7) in terms of intervals.

To this aim, let us consider two intervals $\mathbb{s}$ and $\mathbb{t}$, with $\mathbb{s}, \mathbb{t} \in \Im\Re$. The distance between $\mathbb{s}$ and $\mathbb{t}$ can be defined, following Hausdorff, as

$$d(\mathbb{s}, \mathbb{t}) = \sup\{|\overline{s} - \overline{t}|, |\underline{s} - \underline{t}|\} = |\check{s} - \check{t}| + |\Delta s - \Delta t|, \tag{14}$$

with $(\Im\Re, d)$ a complete metric space.

Given two interval matrices $\mathbb{S}, \mathbb{T} \in \Im\Re^{n\times p}$, the distance matrix between $\mathbb{S}$ and $\mathbb{T}$, $d(\mathbb{S}, \mathbb{T})$, is a matrix whose elements are the component-wise Hausdorff distances between the intervals $\mathbb{s}_{ij}$ and $\mathbb{t}_{ij}$, the elements of $\mathbb{S}$ and $\mathbb{T}$, respectively, [16,17]. The matrix $d(\mathbb{S}, \mathbb{T})$ is nonnegative, and several different matrix norms have been used to synthesize the information of this distance matrix [18,19]. The particular choice of a norm is arbitrary [20] and it is related to the goals of the analysis. In our case, in order to keep both the distances between the midpoints and the ranges simultaneously under control, and in analogy with the archetypal analysis in the classical case, we choose the Frobenius norm of the matrix $d(\mathbb{S}, \mathbb{T})$,

$$\|d(\mathbb{S}, \mathbb{T})\|_F = \left\| \left| \check{\mathbf{S}} - \check{\mathbf{T}} \right| + |\Delta\mathbf{S} - \Delta\mathbf{T}| \right\|_F, \tag{15}$$

which defines a metric on the set of interval matrices.

Given this metric space, for each m, the matrix $\mathbb{A}(m)$ of the m archetypal parallelotopes can be determined by minimizing the distance between the data interval matrix $\mathbb{X}$ and the matrix

$$\tilde{\mathbb{X}}(m) = \mathbf{\Gamma}(m)\mathbb{A}(m) = \mathbf{\Gamma}(m)\mathbf{B}'(m)\mathbb{X}, \quad \tilde{\mathbb{X}}(m) \in \Im\Re^{n\times p}, \tag{16}$$

i.e. the data matrix reconstructed by the m archetypal parallelotopes.

Thus, given m and the quantity:

$$\begin{aligned}\mathbb{RSS}(m) &= \left\| d\left(\mathbb{X}, \tilde{\mathbb{X}}(m)\right) \right\|_F \\ &= \left\| \left| \check{\mathbf{X}} - \mathbf{\Gamma}(m)\mathbf{B}'(m)\check{\mathbf{X}} \right| + \left| \Delta\mathbf{X} - \mathbf{\Gamma}(m)\mathbf{B}'(m)\Delta\mathbf{X} \right| \right\|_F,\end{aligned} \tag{17}$$

the m archetypes solve the minimization problem:

$$\begin{aligned}\min_{\mathbf{\Gamma}(m),\mathbf{B}(m)} \mathbb{RSS}(m) &= \min_{\mathbf{\Gamma}(m),\mathbf{B}(m)} \left\| d\left(\mathbb{X}, \tilde{\mathbb{X}}(m)\right) \right\|_F \\ &= \min_{\mathbf{\Gamma}(m),\mathbf{B}(m)} \left\| \left| \check{\mathbf{X}} - \mathbf{\Gamma}(m)\mathbf{B}'(m)\check{\mathbf{X}} \right| \right. \\ &\quad \left. + \left| \Delta\mathbf{X} - \mathbf{\Gamma}(m)\mathbf{B}'(m)\Delta\mathbf{X} \right| \right\|_F,\end{aligned} \tag{18}$$

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

$$\mathbf{\Gamma}(m)\mathbf{1}_m = \mathbf{1}_n, \quad \mathbf{B}'(m)\mathbf{1}_n = \mathbf{1}_m, \tag{19}$$

where $\mathbf{1}_n$ and $\mathbf{1}_m$ are the all-ones vectors of dimensions n and m, respectively.

The $\mathbb{RSS}(m)$ is a decreasing function of m, and it can be used to decide upon the number of archetypes as in the single-valued case.

5. REPRESENTING AND RECONSTRUCTING DATA THROUGH ARCHETYPES

As can be deduced from the previous discussions, interval archetypal analysis produces three different results: (i) the $m \times p$ archetypes' matrix $\mathbb{A}(m)$, (ii) the $n \times m$ weighting coefficients' matrix $\mathbf{\Gamma}(m)$, and (iii) the $m \times n$ weighting coefficients' matrix $\mathbf{B}(m)$, where $\mathbb{A}(m)$ is an interval data matrix while $\mathbf{\Gamma}(m)$ and $\mathbf{B}(m)$ are classical single-valued matrices.

When the interval archetypes are used in a data analysis perspective, we can distinguish two different phases. In a first step, we start by looking both at the interval archetypes' matrix $\mathbb{A}(m)$, and, separately, at the midpoint and range archetypes' matrices $\check{\mathbf{A}}(m)$ and $\Delta\mathbf{A}(m)$. In this phase, we use archetypes to synthesize data, in order to determine which type of data each archetype represents, to understand archetype characteristics in terms of the original features, and to compare archetypes with respect to the whole data set. With these goals, archetypes can be graphically displayed in any plot of the original data. In these graphical representations, archetypes are interpreted

in terms of distance or proximity with respect to other data points, while bearing in mind that they may or may not coincide with observed statistical units. Note that the interval archetypes can be represented either in $\Im\Re^p$, as paralletopes, through the matrix $\mathbb{A}(m)$, or separately in $\Re^p$, as points in the midpoint and range spaces, through the matrices $\check{\mathbf{A}}(m)$ and $\Delta\mathbf{A}(m)$, respectively. For single-valued data representation, i.e. in midpoint and range spaces, we suggest to use the scatter plot matrices, parallel coordinate plots [21,22], and percentile profile plots [2,23]. While, for interval-valued archetypes, zoom-star plot and scatter plots for interval data can be used [19].

The second step, which is no less important than the first, looks at data through the archetypes by taking into account the coefficients in the $\mathbf{\Gamma}(m)$ matrix. Indeed, the $\boldsymbol{\gamma}'_i(m)$ coefficients play a central role in the analysis and have many interesting interpretations, since they are:

1. weighting coefficients for reconstructing data,
2. barycentric coordinates, and
3. values of a membership function.

1.) $\boldsymbol{\gamma}'_i(m)$ *as weighting coefficients*. From Eq. (16), the $\boldsymbol{\gamma}'_i(m)$ coefficient vectors make it possible to reconstruct each data point $\mathbb{x}_i$ starting from the archetypes. Each interval observation can be reconstructed in two equivalent ways. The first one consists in combining the reconstructions in the midpoint and range spaces:

$$\tilde{\mathbb{x}}'_i(m) = [\check{\tilde{\mathbf{x}}}'_i(m) - \Delta\tilde{\mathbf{x}}'_i(m), \check{\tilde{\mathbf{x}}}'_i(m) + \Delta\tilde{\mathbf{x}}'_i(m)].$$

The second one consists in using directly the weighted sum of the interval archetypes:

$$\tilde{\mathbb{x}}'_i(m) = \boldsymbol{\gamma}'_i(m)\mathbb{A}(m).$$

These two ways are equivalent because the weighting coefficients $\boldsymbol{\gamma}'_i(m)$ are constrained to be the same in the two spaces and, hence, we have that $\check{\tilde{\mathbf{x}}}'_i(m) = \boldsymbol{\gamma}'_i(m)\check{\mathbf{A}}(m)$ and $\Delta\tilde{\mathbf{x}}'_i(m) = \boldsymbol{\gamma}'_i(m)\Delta\mathbf{A}(m)$.

The $\gamma_{ij}(m)$ values can be interpreted as the *contribution* of each archetype $\mathbb{a}'_j(m)$ to a given statistical unit $\mathbb{x}'_i$: if $\gamma_{ij}(m)$ is equal to 1, the statistical unit coincides with the archetype; while if $\gamma_{ij}(m)$ is equal to 0, the archetype does not contribute to reconstruct the statistical unit at all. More generally, each statistical unit is reconstructed thanks to the contribution of several archetypes, proportionally to the $\gamma_{ij}(m)$ values.

2.) $\boldsymbol{\gamma}'_i(m)$ *as barycentric coordinates*. The $\check{\mathbf{a}}'_j(m)$'s are located on the boundary of the convex hull of the midpoints $\check{\mathbf{x}}'_i$, and the $\Delta\mathbf{a}'_j(m)$'s are located on the boundary of convex hull of the ranges $\Delta\mathbf{x}'_i$. Hence, the $\check{\mathbf{a}}'_j(m)$'s and the $\Delta\mathbf{a}'_j(m)$'s are vertices of convex polytopes in the midpoint and ranges spaces, respectively.

In these spaces, for each data point $\mathbb{x}'_i = [\check{\mathbf{x}}'_i - \Delta\mathbf{x}'_i, \check{\mathbf{x}}'_i + \Delta\mathbf{x}'_i]$, new coordinates $(\lambda_{i1}, \ldots, \lambda_{im})$ and $(\mu_{i1}, \ldots, \mu_{im})$ can be obtained by solving the equations:

$$(\lambda_{i1} + \cdots + \lambda_{im})\check{\mathbf{x}}'_i = \lambda_{i1}\,\check{\mathbf{a}}'_1 + \cdots + \lambda_{im}\,\check{\mathbf{a}}'_m, \tag{20}$$

$$(\mu_{i1} + \cdots + \mu_{im})\Delta\mathbf{x}'_i = \mu_{i1}\,\Delta\mathbf{a}'_1 + \cdots + \mu_{im}\,\Delta\mathbf{a}'_m. \tag{21}$$

The coefficients $(\lambda_{i1}, \ldots, \lambda_{im})$ and $(\mu_{i1}, \ldots, \mu_{im})$ are the barycentric coordinates [24] of $\check{\mathbf{x}}'_i$ and $\Delta\tilde{\mathbf{x}}'_i$ in the spaces having $\check{\mathbf{a}}_1, \ldots, \check{\mathbf{a}}_m$ and $\Delta\mathbf{a}_1, \ldots, \Delta\mathbf{a}_m$ as bases, respectively. The archetypes themselves have barycentric coordinates $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0)$, …, $(0, 0, \ldots, 1)$.

The reconstructed data point $\tilde{\mathbb{x}}'_i(m)$ has barycentric coordinates in these associated spaces as well:

$$(\lambda_{i1} + \cdots + \lambda_{im})\check{\tilde{\mathbf{x}}}'_i(m) = \lambda_{i1}\,\check{\mathbf{a}}'_1 + \cdots + \lambda_{im}\,\check{\mathbf{a}}'_m \tag{22}$$

$$(\mu_{i1} + \cdots + \mu_{im})\Delta\tilde{\mathbf{x}}'_i(m) = \mu_{i1}\,\Delta\mathbf{a}'_1 + \cdots + \mu_{im}\,\Delta\mathbf{a}'_m. \tag{23}$$

By Eq. (16), it is easy to show that the $\gamma_{ij}(m)$ coefficients solve both Eqs. (22) and (23), i.e. $\lambda_{ij} = \mu_{ij} = \gamma_{ij}(m)$. Hence, we have that the $\boldsymbol{\gamma}'_i(m)$ coefficient vectors are the barycentric coordinates for the reconstructed points in a common associated space. Such a space, spanned by the archetypes $\mathbb{a}'_j(m)$, is always a space of real points and the archetypes are a non-orthogonal basis of this space [11].

We note that, given the geometric properties of the barycentric coordinates, the data points actually belong to

Table 1. Bats data set [26].

i	Species	Head	Tail	Height	Forearm
1	PIPC	33, 52	26, 33	4, 7	27, 32
2	PRH	35, 43	24, 30	8, 11	34, 41
3	MOUS	38, 50	30, 40	7, 8	32, 37
4	PIPS	43, 48	34, 39	6, 7	31, 38
5	PIPN	44, 48	34, 44	7, 8	31, 36
6	MDAUB	41, 51	30, 39	8, 11	33, 41
7	MNAT	42, 50	32, 43	8, 9	36, 42
8	MDEC	40, 45	39, 44	9, 9	36, 42
9	MGP	45, 53	35, 38	10, 12	39, 44
10	OCOM	41, 51	34, 50	9, 10	34, 50
11	MBEC	46, 53	34, 44	9, 11	39, 44
12	SBOR	48, 54	38, 47	9, 11	37, 42
13	BARB	44, 58	41, 54	6, 8	35, 41
14	OGRIS	47, 53	43, 53	7, 9	37, 41
15	SBIC	50, 63	40, 45	8, 10	40, 47
16	FCHEV	50, 69	30, 43	11, 13	51, 61
17	MSCH	52, 60	50, 60	10, 11	42, 48
18	SCOM	62, 80	46, 57	9, 12	48, 56
19	NOCT	69, 82	41, 59	10, 12	45, 55
20	GMUR	65, 80	48, 60	12, 16	55, 68
21	MGES	82, 87	46, 57	11, 12	58, 63

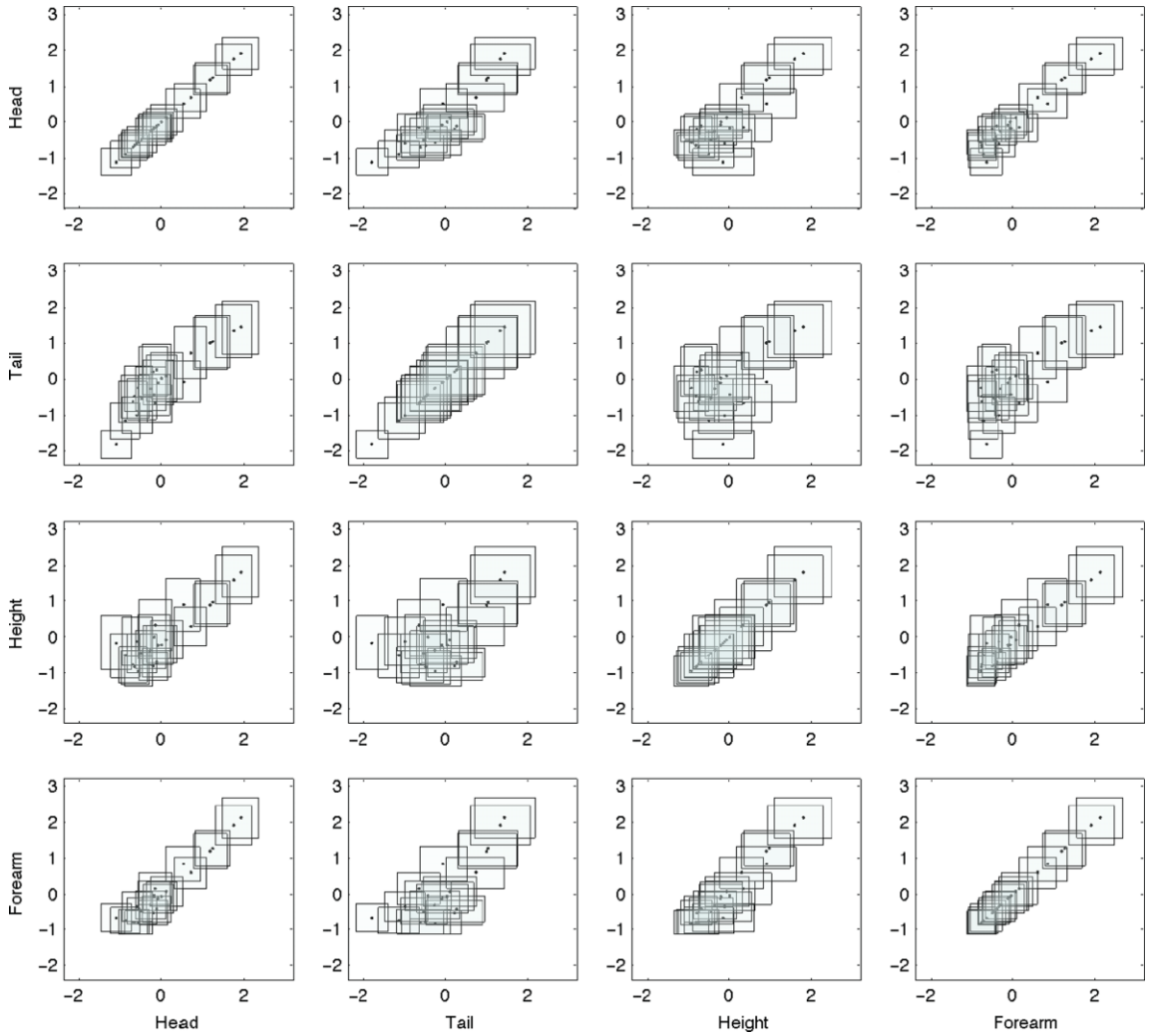

Fig. 1 Scatter plot matrix for Bats data set. The data have been standardized by centering midpoints with respect to their mean $\mu(\check{\mathbf{x}}_j)$ and by scaling by their standard deviation $\sigma(\check{\mathbf{x}}_j)$; ranges have been scaled by dividing by the corresponding midpoints' standard deviation $\sigma(\check{\mathbf{x}}_j)$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

an $(m-1)$ dimensional subspace of the associated space spanned by the archetypes, and the reconstructed data $\tilde{\mathbb{x}}'_i(m)$ are embedded in *a regular hyper-tetrahedron* with unit edges. Furthermore, given the correspondence between each $\mathbb{x}'_i$ and each $\tilde{\mathbb{x}}'_i(m)$ (i.e. $\mathbb{x}'_i$ is equal to $\tilde{\mathbb{x}}'_i(m)$ plus a residual), each original point $\mathbb{x}'_i$ can be represented (through its corresponding $\tilde{\mathbb{x}}'_i(m)$) by the coefficient vectors $\boldsymbol{\gamma}'_i(m)$ in the space spanned by the archetypes. Hence, the $\boldsymbol{\gamma}'_i(m)$ may be exploited to map the original data into this lower dimensional space.

In such a space, data can be graphically analyzed and represented, for example, through a parallel coordinate plot [21,22] in which the axes correspond to the archetypes. Any graphical representation in this space provides an outward–inward perspective on the data and allows us to highlight many data patterns such as small groups, gaps in the data structure, outlying values, asymmetries and irregularities in the data shape.

3.) $\boldsymbol{\gamma}'_i(m)$ *as membership values*. Given the properties exemplified in point 2, the $\boldsymbol{\gamma}'_i(m)$ coefficients reflect the relative proximity of the observations to each archetype and can be properly interpreted as values of a membership function. They can be used to cluster interval-valued data in a very simple way [25]. This can be achieved by choosing a threshold value γ^* for the membership function, and by assigning each interval data $\mathbb{x}'_i$ to the cluster around the j^{th} archetype $\mathbb{a}'_j(m)$ if $\gamma_{ij}(m) > \gamma^*$.

Concerning the $\mathbf{B}(m)$ matrix, even though its values could be interpreted as the *contribution* of each interval observation to each archetype, we believe that it may not be of much interest for the analysis.

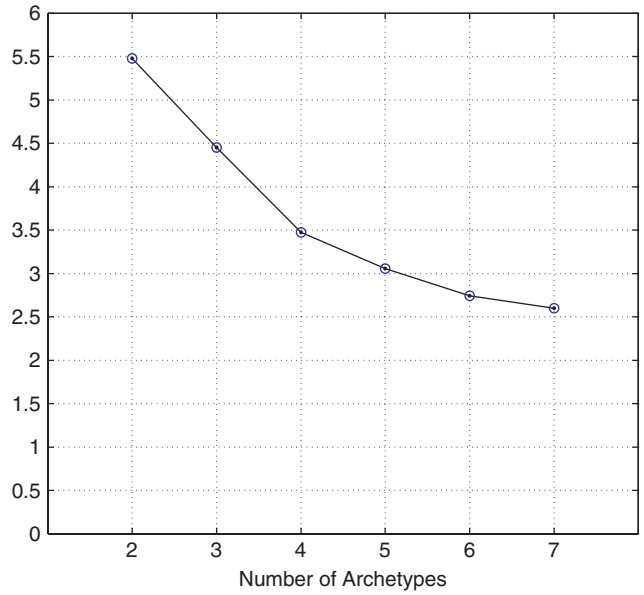

Fig. 2 Bats data set: $\mathbb{RSS}(m)$ function, $m = 2, \ldots, 7$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

6. INTERVAL ARCHETYPES IN ACTION: A REAL DATA EXAMPLE

To provide a clear illustration of how interval archetypes can be a powerful tool for the exploratory analysis of interval data, we use the data set related to characteristics of bats [26]. For this data set, a key question consists of comparing different species of bats with each other in order to highlight the similarities or anomalies. With this aim, principal component analysis for interval data has been applied by Billiard *et al.* [26]. The tools proposed in this paper allow the analyst not only to make a comparison and an exploration of the data, but also to find some species representing all the others.

6.1. The Bats Data Set

The data set (Table 1) has already been presented in Billard *et al.* [26] and refer to 21 species of bats (PIPC,. . ., MGES); species identifiers are abbreviations of the longer biological Latin descriptor, e.g. 'BARB' is the species *Barbastella barbastellus*. Four variables have been measured on each bat: Head = head size, Tail = tail length, Height = height, and Forearm = forearm length.

In the following, the data have been standardized according to the usual procedure for interval data: midpoints $\check{\mathbf{x}}_j$ have been centered with respect to their own mean $\mu(\check{\mathbf{x}}_j)$, with $\mu(\check{\mathbf{x}}_j) = \frac{\sum_{i=1}^{n} \check{x}_{ij}}{n}$, and scaled by their standard deviation $\sigma(\check{\mathbf{x}}_j)$, with $\sigma(\check{\mathbf{x}}_j) = \sqrt{\frac{\sum_{i=1}^{n} [\check{x}_{ij} - \mu(\check{\mathbf{x}}_j)]^2}{n}}$; ranges have been scaled by dividing by the corresponding midpoints' standard deviation $\sigma(\check{\mathbf{x}}_j)$ [27].

An interval data matrix can be represented through a scatter plot matrix (SPLOM) for interval data. In such a SPLOM, unlike the case for single-valued data, scatter plots on the diagonal have an interpretation. They provide information on the correlation between midpoints and ranges for each variable [28]. Indeed, as the square size depends on the range and the square location depends on the midpoint, by looking at the variation of square sizes with their locations, we can graphically appreciate if there is a dependence between ranges and midpoints for each variable. The absence of high correlations between midpoints and ranges implies that they express different sources of variability in the data set.

Figure 1 exhibits the SPLOM for the data at hand. The scatter plots on the diagonal do not show evidence of dependence patterns between midpoints and ranges, as confirmed by the low Pearson linear correlation indexes (the highest value is equal to 0.466, corresponding to the Tail variable). The remaining off-diagonal displays in the SPLOM reveal a general positive correlation between all pairs of variables.

6.2. Interval Archetypes for Data Synthesis and Comparison

This section presents the solution results for $m = 4$ archetypes. The choice of the number of archetypes is based on the behavior of the $\mathbb{RSS}(m)$ function (Fig. 2). For $m = 2, 3, 4$, i.e. for 2, 3, and 4 archetypes, the $\mathbb{RSS}(m)$ function shows a sharp decrease, before becoming flatter for $m > 4$. Hence, a number of archetypes greater than 4

Table 2. Coordinates of interval archetypes for $m = 4$. The data have been standardized by centering midpoints with respect to their mean $\mu(\check{\mathbf{x}}_j)$ and by scaling by their standard deviation $\sigma(\check{\mathbf{x}}_j)$; ranges have been scaled by dividing by the corresponding midpoints' standard deviation $\sigma(\check{\mathbf{x}}_j)$.

$\mathbb{A}(4)$	Head	Tail	Height	Forearm
$\mathrm{a}_1'(4)$	1.584, 2.413	0.672, 2.136	1.058, 2.342	1.585, 2.639
$\mathrm{a}_2'(4)$	0.315, 0.359	−0.660, 1.984	−0.365, 0.331	−0.327, 0.303
$\mathrm{a}_3'(4)$	−1.462, −0.790	−2.235, −1.447	−0.658, 0.824	−0.930, −0.118
$\mathrm{a}_4'(4)$	−1.540, −0.169	−1.854, −1.002	−2.551, −1.200	−1.721, −1.114

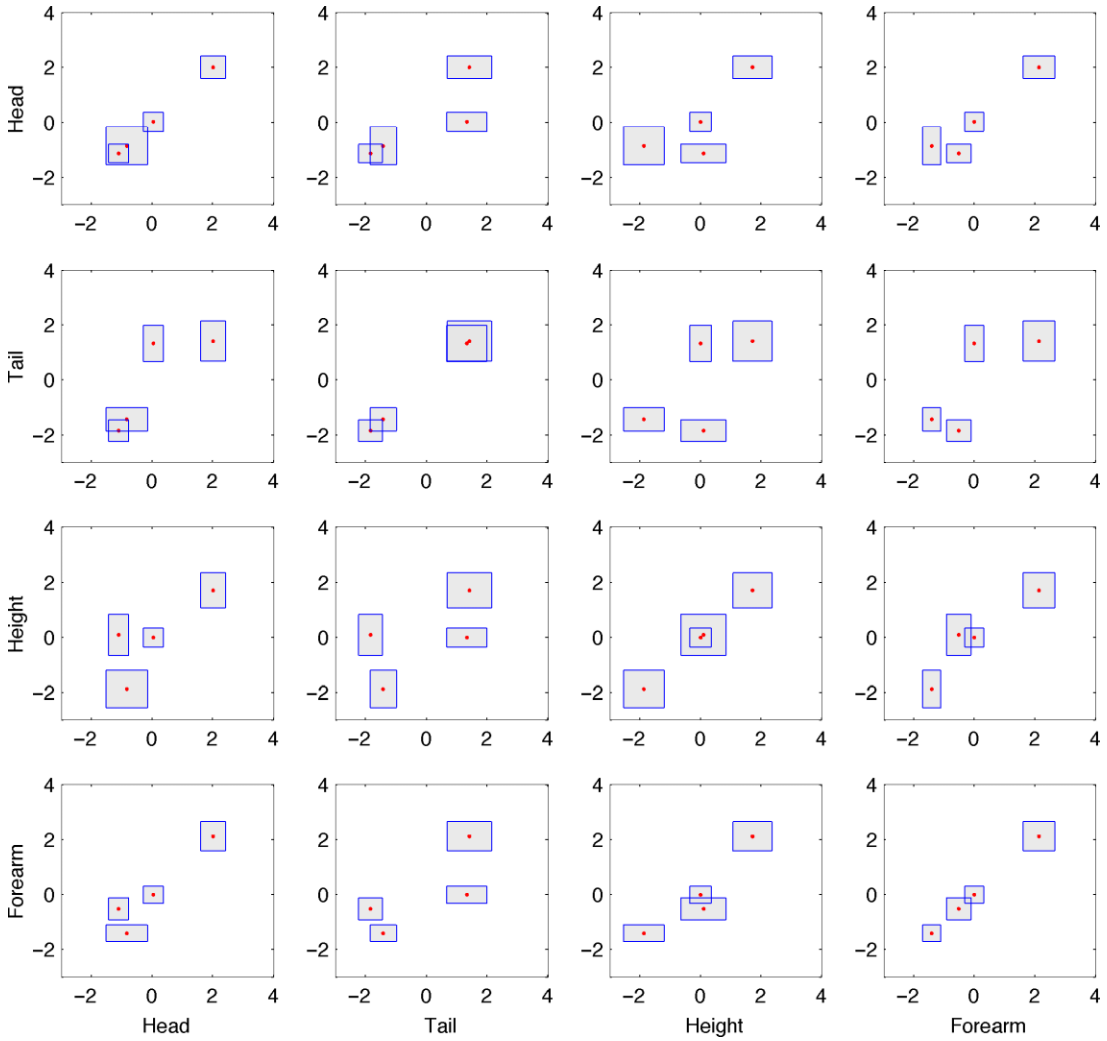

Fig. 3 SPLOM of four interval archetypes in the original space. The data have been standardized by centering midpoints with respect to their mean $\mu(\check{\mathbf{x}}_j)$ and by scaling by their standard deviation $\sigma(\check{\mathbf{x}}_j)$; ranges have been scaled by dividing by the corresponding midpoints' standard deviation $\sigma(\check{\mathbf{x}}_j)$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

do not yield a relevant improvement in the description of the data through the archetypes.

As discussed in Section 5, two different phases of the analysis can be envisioned. In the first, the aim is to understand which kind of observed unit each archetype synthesizes in terms of the original features. Table 2 exhibits the interval archetypes for $m = 4$. The archetypes are visualized in the original space of the standardized data through a SPLOM for interval-valued data (Fig. 3). Moreover, by using the midpoint-range notation, we represent the midpoint archetypes as points in a SPLOM, along with the midpoints of the whole data set (Fig. 4), and through the percentile profile plots (Fig. 5). By looking at both Table 2 and Figs 3, 4, and 5, it can be noticed that the archetypes $\mathbb{a}'_1(4)$ and $\mathbb{a}'_4(4)$ represent the two extreme individuals, the biggest and the smallest, respectively, while archetypes $\mathbb{a}'_2(4)$ and $\mathbb{a}'_3(4)$, even if they are located on the boundary of the data scatter, represent species with *average* profiles. In particular, from the percentile profile plot in Fig. 5, $\mathbb{a}'_2(4)$ represents an archetype for species characterized by median values of all the variables except the height, while the third archetype $\mathbb{a}'_3(4)$ is characterized by small head and tail and median height and forearm. As for $\mathbb{a}'_1(4)$ we note that it is approximately at the 90th percentile for all the characteristics, while $\mathbb{a}'_4(4)$ is approximately at the 15th percentile for the all features.

In the second step, in order to look at the data through the archetypes and to compare the latter with respect to the other data, we look at the $\mathbf{\Gamma}(m)$ matrix defined in Eq. (16). For the Bats data set, the $\mathbf{\Gamma}(4)$ matrix is reported in Table 3.

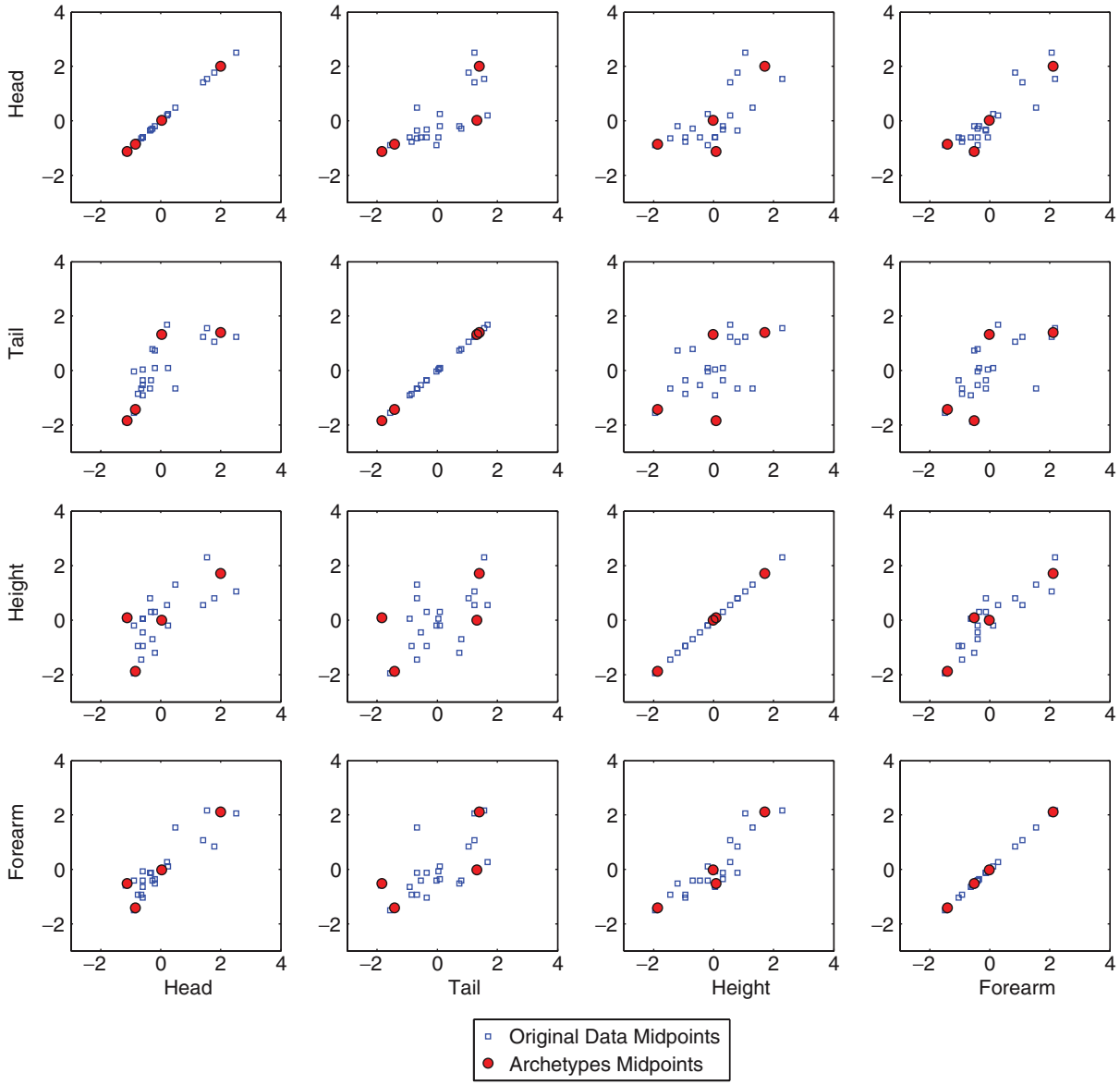

Fig. 4 SPLOM of midpoints of four interval archetypes (circles) and midpoints of the original data (cross). The data have been standardized by centering midpoints with respect to their mean $\mu(\check{\mathbf{x}}_j)$ and by scaling by their standard deviation $\sigma(\check{\mathbf{x}}_j)$; ranges have been scaled by dividing by the corresponding midpoints' standard deviation $\sigma(\check{\mathbf{x}}_j)$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 3 can be read by rows and by columns. The rows can be interpreted as the *contribution* of each archetype to the reconstruction of a given statistical unit ($\boldsymbol{\gamma}_i'(m)$ *as weighting coefficients*, see Section 5). If one of the $\gamma_{ij}(4)$ values is equal to 1, then the statistical unit corresponds to the j^{th} archetype. For the data set we are analyzing, the statistical units 1 and 20 correspond to archetypes $\mathbb{a}_4'(4)$ and $\mathbb{a}_1'(4)$, respectively. This implies that to reconstruct each one of these units only one archetype is needed. On the contrary, all the other observations are reconstructed through all the four archetypes. For example, by looking at the coefficients on row 15, we note that the 'SBIC' observation is expressed as a weighted sum of all four archetypes, i.e. it represents a sort of multivariate average species.

Figure 6 illustrates these properties by considering the cases of units 1 and 15. Each unit is represented through a panel of four scatter plots, one for each archetype. In the scatter plots, the axes coincide with the variables Height and Forearm (selected from the four available for the sake of illustration). Each plot in the panel graphically represents the part of the statistical unit reconstructed by archetypes.

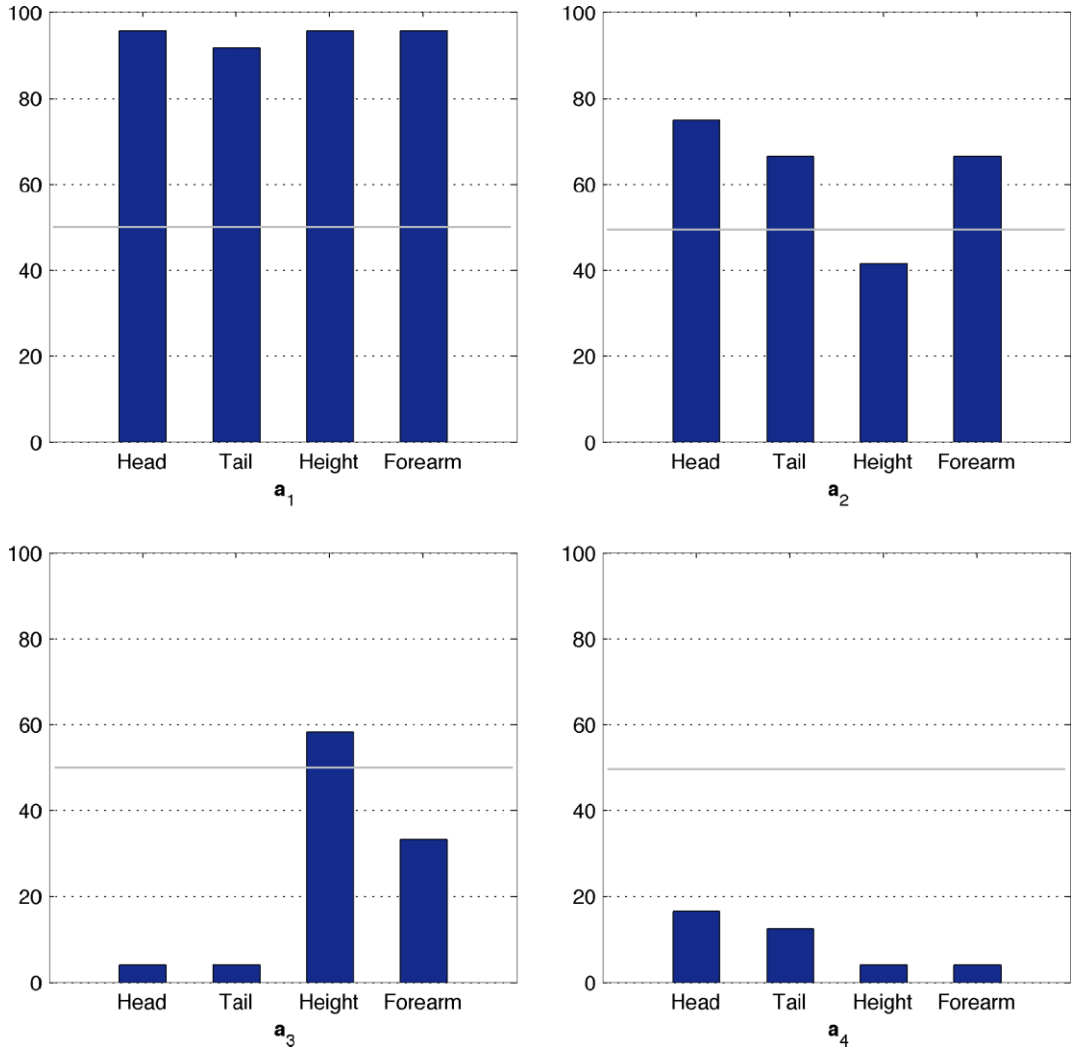

Fig. 5 Bats data set. Percentile profile plots of the midpoints of the four interval archetypes $\mathbb{a}'_1, \ldots, \mathbb{a}'_4$, left to right top to bottom. The vertical axes represent the percentage cumulative distribution. The 50% line is superimposed. Each archetype $\breve{\mathbf{a}}_j(4)$ is represented by a sequence of vertical bars, one for each variable. The height of k-th bar corresponds to the percentile of the archetype in the empirical distribution function of the data plus the archetypes in the midpoint space. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In the panel from left to right and from top to bottom, the contribution of each archetype is successively added. In the fourth plot (bottom-right), the complete reconstruction is represented. The left panel (Fig. 6*a*) refers to unit 1 'PIPC'. As it coincides with the fourth archetype, the contribution to reconstruction appears only in the bottom-right plot. On the contrary, in the right panel (Fig. 6*b*) corresponding to unit 15 'SBIC', it is evident that the contribution of all four archetypes is necessary to reconstruct the unit.

The $\mathbf{\Gamma}(4)$ matrix in Table 3 can be read also by columns, and the $\gamma_{ij(m)}$ values can be interpreted *as the barycentric coordinates* of the data points in the space spanned by the four archetypes or *as a membership degree* of each data to a given archetype (Section 5). With four archetypes, the space spanned by the archetypes is a four-dimensional one. In each column, the values close to 1 indicate the data points close or equal to an archetype. In Table 3, for each column the largest value is in boldface. In particular, note that, as $\gamma_{20,1} = 1$, the observation numbered 20 (species 'GMUR') coincides with $\mathbb{a}'_1(4)$; then this species can be assumed to represent the archetype of big bats. On the contrary, as $\gamma_{1,4} = 1$, the observation numbered 1 (species 'PIPC') coincides with $\mathbb{a}'_4(4)$. Then this species represents the archetype of small bats. Furthermore, the observation numbered 2 (species 'PRH') almost coincides with $\mathbb{a}'_3(4)$ as $\gamma_{2,3}$ is close to 1, and the observation numbered 17 (species 'MSCH') is very close to $\mathbb{a}'_2(4)$. In order to cluster the $\mathbb{x}'_i$'s in four clusters $(C_1, \ldots, C_4)$ around seeds coinciding

Table 3. Bats data set: weighting coefficient matrix $\mathbf{\Gamma}(m)$ for $m = 4$.

i	Species	$\gamma_1(4)$	$\gamma_2(4)$	$\gamma_3(4)$	$\gamma_4(4)$
1	**PIPC**	0	0	0	**1.000**
2	**PRH**	0.000	0.000	**0.983**	0.017
3	MOUS	0.000	0.243	0.242	0.515
4	PIPS	0.000	0.262	0.000	0.738
5	PIPN	0.000	0.396	0.096	0.508
6	MDAUB	0.039	0.240	0.643	0.078
7	MNAT	0.008	0.368	0.363	0.261
8	MDEC	0.000	0.556	0.327	0.117
9	MGP	0.184	0.182	0.634	0.000
10	OCOM	0.112	0.466	0.336	0.086
11	MBEC	0.113	0.355	0.532	0.000
12	SBOR	0.040	0.589	0.371	0.000
13	BARB	0.000	0.689	0.000	0.311
14	OGRIS	0.000	0.726	0.000	0.274
15	SBIC	0.318	0.234	0.052	0.396
16	FCHEV	0.516	0.000	0.484	0.000
17	**MSCH**	0.173	**0.827**	0.000	0.000
18	SCOM	0.686	0.153	0.000	0.160
19	NOCT	0.668	0.212	0.000	0.120
20	**GMUR**	**1.000**	0	0	0
21	MGES	0.965	0.000	0.000	0.035

Notes: boldface the largest value for each column.

with the archetypes, we set a threshold γ^* equal to 0.5, i.e. we assign a data $\mathbb{x}_i'$ to a cluster C_j if $\gamma_{ij} > \gamma^*$. By this rule ($\gamma_{ij} > 0.5 \Rightarrow \mathbb{x}_i' \in C_j$), for example, the observations numbered 16, 18, 19, 20, and 21 belong to the cluster C_1 around the first archetype $\mathbb{a}_1'(4)$, and represent the cluster of big bat species. Note that the observations numbered 7, 10, and 15 do not clearly belong to any cluster.

To visualize the observations in the four-dimensional space spanned by the archetypes, we note that all the reconstructed points actually belong to a three-dimensional subspace, more precisely a regular tetrahedron with unit edges. The γ_{ij} coefficients map the original data in this tetrahedron. Figure 7 shows how, in cases like the present dataset (i.e. with a number of archetypes $m \leq 4$), it is easy to look at data in the space spanned by the archetypes through graphically dynamic tools such as a tourplot [29–32]. In general, for any $m \geq 3$, it could be convenient to use the parallel coordinate plot. By using the four archetypes as axes of parallel coordinate plots, in Fig. 8*a* (left panel) we display all the data highlighting, through different line styles, the previously identified four groups. To understand the cluster structures better, in Fig. 8*b* (right panel), we represent each group separately.

By looking at Figs 7 and 8, we see that all the previous results are confirmed but they are more clearly highlighted. In addition, the plots also reveal some details not evident from Table 3. In particular, we note that the groups C_1 and C_2 are less homogeneous than the other two. Indeed, one datum in the first group (unit 16) presents a slightly different behavior from the others, while in the second group there are three data that are very similar to each other and two data (units 8 and 12) that present higher coordinates on the third archetype.

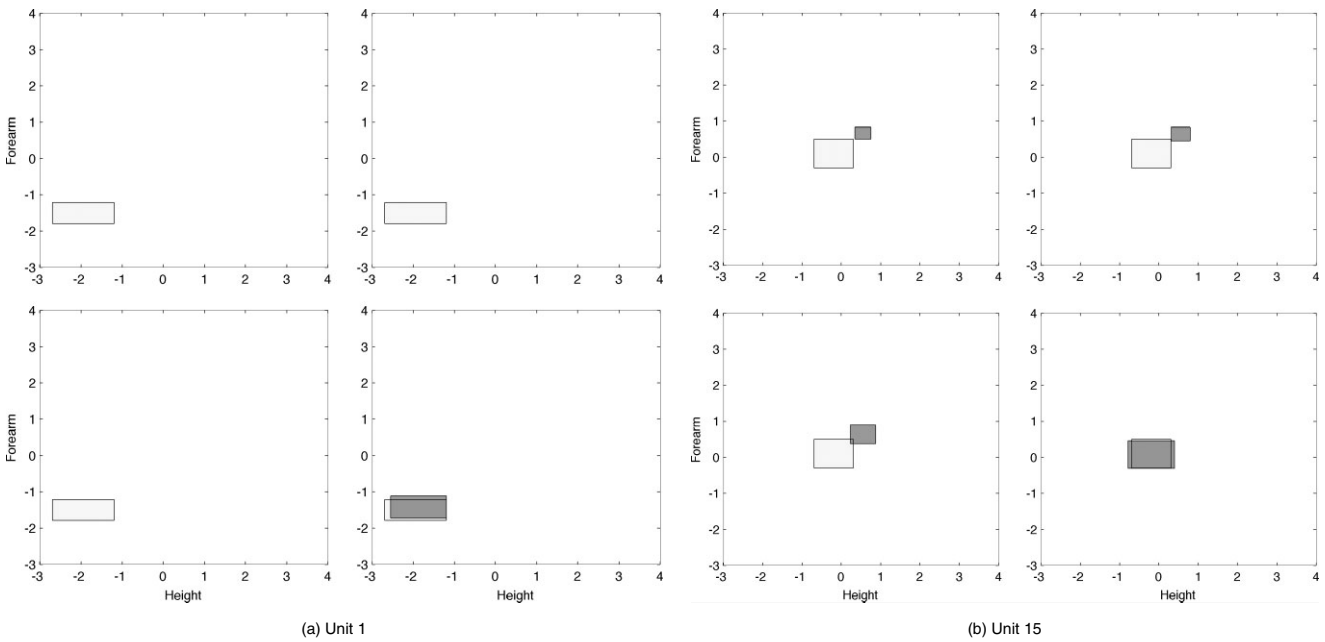

Fig. 6 Reconstruction of observations numbered 1 (*a*) and 15 (*b*) with respect to the variables Height and Forearm. In each panel, plots represent the part of the statistical unit reconstructed by archetypes: in each panel from left to right and from top to bottom, the contribution of each archetype is successively added. In the fourth plot (bottom-right) the complete reconstruction is represented.

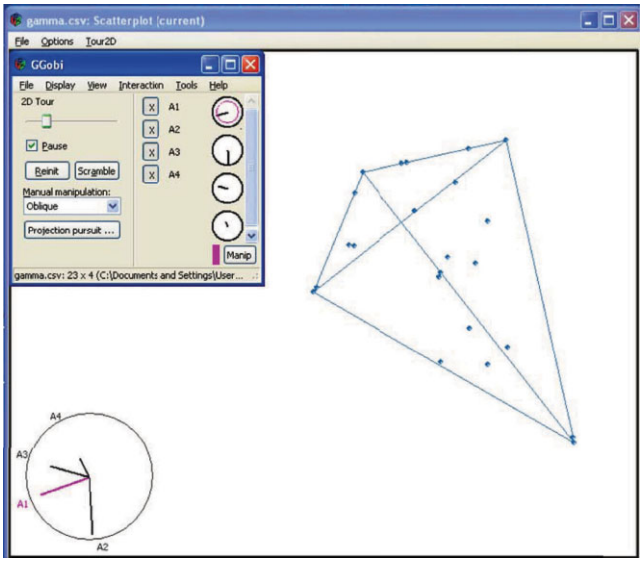

Fig. 7 Tourplot representation of all data in the space spanned by the four archetypes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

6.3. Computational Issues

All results illustrated in this paper have been obtained using the Corsaro and Marino procedure [33], which solves the minimization problem adopting a sequential quadratic programming method. The procedure has been implemented in MATLAB© with the `fmincon` (find minimum of constrained nonlinear multivariable function) MATLAB© routine and the following optimization parameters:

- Maximum number of function evaluations allowed = 1 000 000
- Maximum number of iterations allowed = 100 000
- Tolerances (function value and constraints violation) = $10e^{-5}$.

Here, we consider the problems of the solution stability and of the computational cost to achieve solutions.

As regards the stability issue, we analyze the algorithm performances with respect to the presence of local minima, that may be due to the non-convexity of the objective function [33]. For each value of m, we ran 100 independent replications with random starting solutions and different initial seeds using algorithm in ref. 33. Figure 9 presents the distributions of the 100 values of the $\mathbb{RSS}(m)$ function for $m = 2, \ldots, 7$; each solution distribution is represented through a box plot. It is worth noticing that the solution stability worsens as m increases. Indeed, the variability of the solutions over the runs, highlighted in the boxplots, for $m = 6, 7$ increases and some heavy outlying solutions appear in the distributions.

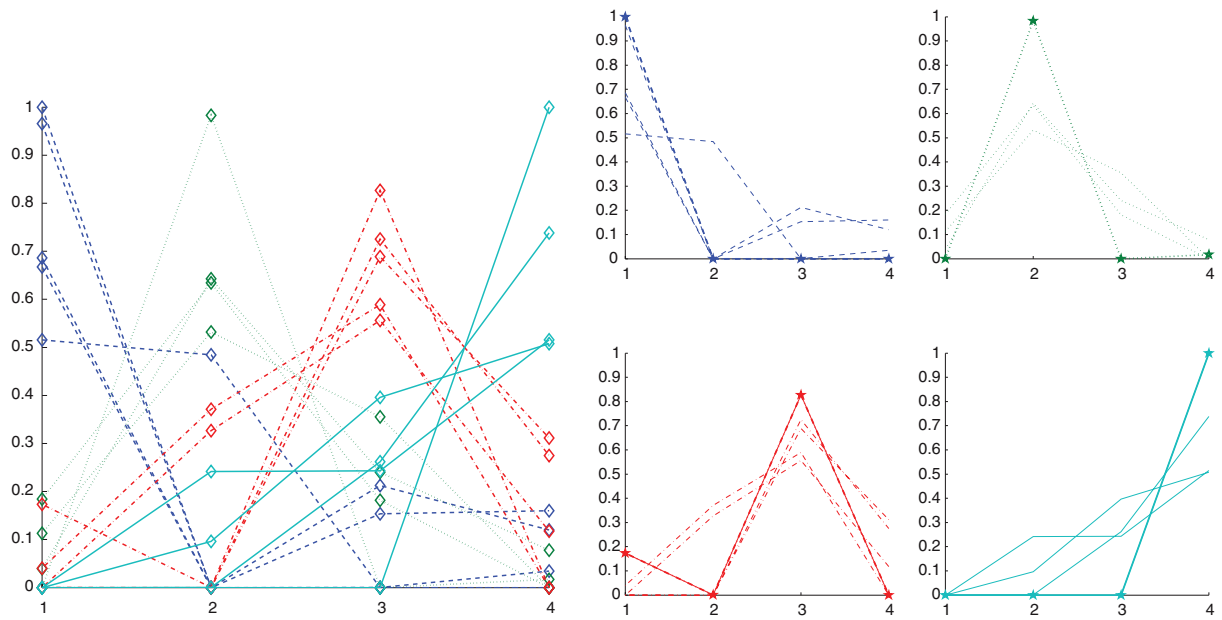

Fig. 8 Bats data set: Parallel coordinate plot of data in the space spanned by the four interval archetypes: (a) the four groups are highlighted through different line styles; (b) each group is represented separately. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 4. Computational cost. CPU average time and maximum value in milliseconds on Intel Mobile Centrino CPU based notebook with 1.66 GHz clock.

Number of archetypes	Average time	Maximum time	Order
2	7.66	50	$\leq 10^1$
3	30.03	58	$\leq 10^2$
4	90.27	183	$\leq 10^3$
5	253.26	386	$\leq 10^3$
6	578.39	1199	$\leq 10^4$
7	1211.76	5186	$\leq 10^4$
8	6068,45	27669	$\leq 10^5$

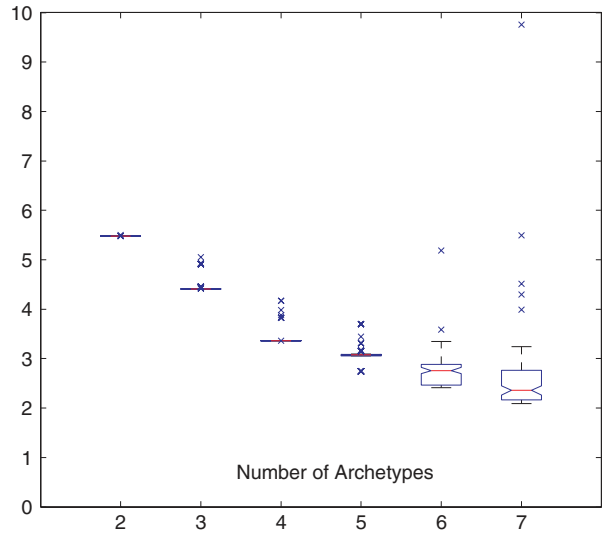

Fig. 9 Parallel boxplots of the $\mathbb{RSS}(m)$ function for different values of m, $m = 2, \ldots, 7$. Each boxplot refers to the $\mathbb{RSS}(m)$ values in 100 replications. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Focusing on the case of $m = 4$, which corresponds to the number of archetypes used in the example illustrated in the paper, 79 times out of 100, the procedure found the minimum $\mathbb{RSS}(4)$ in the interval [3.3628–3.3632], which can be assumed as the global minimum value. In the remaining 21 iterations, the procedure stopped for values in the range [3.3632–4.1720]. These results are extremely encouraging in terms of stability of the solution we used.

With respect to the computational costs, it seems that the procedure is feasible in terms of time up to six archetypes. The CPU time was calculated using the `tic tac` functions available in MATLAB©. The program was run on an Intel Mobile Centrino CPU based notebook with 1.66 GHz clock. Values in Table 4 show that the computing time exponentially increases with m. For $m = 4$, the average CPU time is 0.009 s which becomes 0.025 s for $m = 5$. When m is set to 8, the CPU time is 6.06 s, which corresponds to approximately 3 h of total computational time.

7. CONCLUSION

In this paper, we have proposed the extension of the idea of archetypes in the framework of interval-valued data analysis. Interval-valued data may represent the output of queries to large or huge databases, where groups of homogeneous statistical units are summarized by intervals defined by the minimum and maximum value of each variable. Within this framework, tools are needed to synthesize large amounts of complex data and to represent them through interesting observations. We have shown that interval archetypes may serve this purpose, as they synthesize interval data through linear combinations and represent, therefore, a type of prototypical observation. Moreover, interval archetypes can provide a mean to look at data from an outward–inward point of view and, hence, they can be a powerful tool for the exploratory analysis of interval data. In this paper, we have proposed a formal definition of interval archetypes along with a discussion of their statistical and geometrical properties. We have also suggested a procedure to use them in practice. Open issues that could be tackled regard the possibility of obtaining a more refined clustering procedure starting from the archetypes and how to extend the idea of archetypes to other types of symbolic data.

ACKNOWLEDGMENTS

We are grateful to our anonymous referees and to the editor for their valuable comments.

REFERENCES

[1] L. Billard, Some analyses of interval data, J Comput Inf Technol 4 (2008), 225–233.
[2] A. Cutler and L. Breiman, Archetypal analysis, Technometrics 36 (1994), 338–347.
[3] E. Stone and A. Cutler, Introduction to archetypal analysis of spatio-temporal dynamics, Physica D 96(1-4) (1996), 110–131.
[4] E. Stone, Exploring archetypal dynamics of pattern formation in cellular flames, Physica D 161(3-4) (2002), 163–186.
[5] B. H. P. Chan, D. Mitchell, and L. Cram, Archetypal analysis of galaxy spectra, Mon Not R Astron Soc 338 (2003), 790–795.
[6] S. Marinetti, L. Finesso, and E. Marsilio, Matrix factorization methods: application to thermal NTD/E, NDT E Int 39(8) (2006), 611–616.
[7] S. Marinetti, L. Finesso, and E. Marsilio, Archetypes and principal components of an IR image sequence, Infrared Phys Technol 49(3) (2007), 272–276.
[8] L. Morris and R. Schmolze, Consumer archetypes: a new approach to developing consumer understanding frameworks, J Mark Res 46 (2006), 289–300.

[9] A. Elder and J. Pinnel, Archetypal analysis: an alternative approach to finding defining segments, In 2003 Sawtooth Software Conference Proceedings, Sequim, WA, 2003, 113–129.
[10] S. Li, P. Wang, J. Louviere, and R. Carson, Archetypal analysis: a new way to segment markets based on extreme individuals, In A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution, ANZMAC 2003 Conference Proceedings, Adelaide, 2003.
[11] G. C. Porzio, G. Ragozini, and D. Vistocco, On the use of archetypes as benchmarks, Appl Stochastic Models Bus Ind 25(5) (2008), 419–437.
[12] W. Heavlin, Archetypal analysis of computer performance, In 38th Symposium on the INTERFACE on Massive Data Sets and Stream, Pasadena, CA, 2007.
[13] G. Ragozini and M. R. D'Esposito, A new r-ordering procedure to rank multivariate performances, Quad Stat 10 (2008), 5–21.
[14] H. Bock and E. Diday, Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Heidelberg, Springer Verlag, 2000.
[15] M. R. D'Esposito, F. Palumbo, and G. Ragozini, Archetypal symbolic objects, In 45th Scientific Meeting of the Italian Statistical Society Conference Proceedings, Padua, Italy, 2010.
[16] R. Kearfott, Rigorous Global Search: Continuous Problems, Dordrecht, Netherlands, Kluwer Academic Publishers, 1996.
[17] T. Hickey, Q. Ju, and M. H. Van Emden, Interval arithmetic: from principles to implementation, J ACM 48, (2001), 1038–1068.
[18] M. Chavent, A Hausdorff distance between hyper-rectangles for clustering interval data, In Classification, Clustering, and Data Mining Applications; Studies in Classification, Data Analysis, and Knowledge Organization, Series. D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul, eds. Berlin Heidelberg, Springer, 2004, 333–340.
[19] M. R. D'Esposito, F. Palumbo, and G. Ragozini, Archetypal analysis for interval data in marketing research, Stat Appl, Ital J Appl Stat 18 (2006), 343–358.
[20] E. Hansen, Interval arithmetic in matrix computations, part I, J SIAM: Ser B, Numer Anal 2(2) (1965), 308–320.
[21] A. Inselberg, The plane with parallel coordinates, Vis Comput 1(2) (1985), 69–91.
[22] E. J. Wegman, Hyperdimensional data analysis using parallel coordinates, J Am Stat Assoc 85(411) (1990), 664–675.
[23] M. R. D'Esposito, G. Ragozini, and D. Vistocco, Exploring data through archetypes, In Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization Series, H. Locarek–Junge and C. Weihs Claus, eds. Berlin-Heidelberg, Springer, 2010, 287–298.
[24] H. S. M. Coxeter, Introduction to Geometry, chap. 13.7, Barycentric coordinates (2nd ed.), New York, Wiley, 1969, 216–221.
[25] M. R. D'Esposito, F. Palumbo, G. Ragozini, Archetypal analysis for prototype identification, In GfKl-Cladag 2010 Proceedings, Firenze, Italy, 2010.
[26] L. Billard, A. Douzal-Chouakria, and E. Diday, Symbolic principal component for interval-valued observations, 2009, http://hal.archives-ouvertes.fr/docs/00/36/10/53/PDF/DouzalPCA.pdf [Last accessed November 2011].
[27] P. Giordani and H. A. L. Kiers, Principal component analysis of symmetric fuzzy data, Comput Stat Data Anal, 45(3) (2004), 519–548.
[28] F. Palumbo and C. Lauro, A PCA for interval valued data based on midpoints and radii, In New Development in Psychometric, H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. Meulman, eds. Tokyo, Springer, 2003, 641–648.
[29] D. Cook and D. Swayne, Interactive and Dynamic Graphics for Data Analysis: With R and GGobi, User R, New York, Springer, 2007.
[30] D. Temple Lang, D. Swayne, H. Wickham, and M. Lawrence, RGGobi: Interface between R and GGobi. R package version 2.1.10, 2008.
[31] F. Young, R. Faldowski, and D. Harris, The spreadplot: a graphical spreadsheet of algebraic linked of dynamic plot, In ASA Proceedings Section on Statistical Graphics, Alexandria, VA, American Statistical Association, 1992.
[32] F. W. Young, R. A. Faldowski, and M. M. McFarlane, Multivariate statistical visualization, In Computational Statistics, Handbook of Statistics, Vol . 9, chap. 27, Amsterdam, Elsevier B.V., 1993, 959–998.
[33] S. Corsaro and M. Marino, Archetypal analysis of interval data, Reliab Comput 14 (2010), 105–116.