Principal Component Analysis for Interval-Valued Observations

Ahlame Douzal-Chouakria, Lynne Billard, Edwin Diday

Principal Component Analysis for Interval-Valued Observations

A. Douzal-Chouakria[1]*, L. Billard[2] and E. Diday[3]

[1]*University of Joseph Fourier, Grenoble 1, 38041 Grenoble, Cedex 9, France*

[2]*Department of Statistics, University of Georgia, Athens, GA 30602, USA*

[3]*Ceremade, University of Paris Dauphine, 75775 Paris, Cedex 16, France*

Received 18 January 2010; revised 27 January 2011; accepted 28 January 2011
DOI:10.1002/sam.10118
Published online 8 March 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: One feature of contemporary datasets is that instead of the single point value in the p-dimensional space $\Re^p$ seen in classical data, the data may take interval values thus producing hypercubes in $\Re^p$. This paper studies the vertices principal components methodology for interval-valued data; and provides enhancements to allow for so-called 'trivial' intervals, and generalized weight functions. It also introduces the concept of vertex contributions to the underlying principal components, a concept not possible for classical data, but one which provides a visualization method that further aids in the interpretation of the methodology. The method is illustrated in a dataset using measurements of facial characteristics obtained from a study of face recognition patterns for surveillance purposes. A comparison with analyses in which classical surrogates replace the intervals, shows how the symbolic analysis gives more informative conclusions. A second example illustrates how the method can be applied even when the number of parameters exceeds the number of observations, as well as how uncertainty data can be accommodated. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 229–246, 2011

Keywords: vertices principal components; vertex contributions; correlations; inertia

1. INTRODUCTION

Principal component analysis is a well established method designed to reduce the dimensionality p of a dataset into one of dimension $s \ll p$, so as to facilitate the visualization and extraction of the main trends in a high-dimensional dataset. These techniques have focused on classical datasets whereby each observation is a single point in the p-dimensional space $\Re^p$ [1]. The goal of this paper is to review and compare principal component methodology for interval-valued symbolic data; and to add enhancements to the so-called vertices method.

Interval-valued data can result from aggregation of a typically large dataset into one of more manageable size or one whose focus is on some specific aspect. The nature of the aggregation would vary depending on the scientific questions of interest. For example, in the faces dataset (considered in Section 5), facial characteristics were determined from a series of images by measuring the number of pixels for each image. Aggregating these over the complete set of images produced intervals since understandably different images would contain differing numbers of pixels. Interval-valued data can also arise in their own right such as species. For example, the bat species *Pipistrelle Commune* has height from 4 to 7 mm (i.e., the interval [4,7]) but a particular bat may have a height of 4.3 mm. The list of naturally arising interval data is endless.

In a different direction, some measurements carry an inherent degree of uncertainty and/or imprecision. For example, your assessment of the merits of some entity (e.g., wine quality) can be along the lines of $90 \pm \delta$ with $\delta = 5$ when reasonably sure and $\delta = 10$ when the uncertainty increases. Rather than uncertainty, in order to protect confidentialities, an actual observation of 24 say may be recorded as $(24 - \delta_1,\ 24 + \delta_2)$ for arbitrary δ_1, δ_2 values. Also, we use such notions on a regular basis when we measure, for example, pulse rate as 64 ± 1, that is, [63, 65]. Note however that pulse rates of 64 ± 1 and 64 ± 3, while having the same midpoint have different internal variations,

Correspondence to: A. Douzal-Chouakria (Ahlame.Douzal@imag.fr)

and so are differently valued observations. Any analysis therefore must take into account these internal variations inherent to symbolic data along with the usual external variations familiar to us as between (classical) observations, that is, variance. A review of symbolic data can be found in refs. [2,3].

The problem of reducing a large number of random variables p to a smaller number of principal components $s \ll p$ remains regardless of how the intervals were formed. Therefore, in Section 2, the vertices method for performing a principal component analysis on interval-valued data is presented. A brief summary of this method was given in Ref. [4]. In this work, we complete some of those details and extend the method further. Thus, we make allowance for intervals to be 'trivial' as can happen when a given variable assumes a classical rather than an interval value; that is, the data can be hypercubes in $p' \leq p$-dimensions. General weight functions are introduced in Section 2.2, and calculation of the underlying variance–covariance matrix along with practical computational complexity considerations are presented in Section 2.3. The method is based on extending the methodology for classical data, and we show (in Section 2.4) how the basic classical theory carries through to interval-valued data. Further, we show how the method allows for completely classical data as a special case.

Then, we introduce the concept of vertex contributions to the principal components, a concept not possible in a classical analysis. Unlike classical data which consist of single points in p-dimensional space, interval data consist of hypercubes each consisting of a cloud of vertices. Therefore, the contribution of an observation to the principal component can be broken down into contributions of each vertex. The visualization and hence interpretation of the vertices principal components can therefore be further enhanced by focusing on those vertices whose contributions exceed preassigned bounds (see Section 3).

Other efforts to address interval data include the methods of Lauro and Palumbo [5]. In an attempt to improve the factorial visualization of these analyses, they considered three variants, each based on the interval midpoints (a special case of the centers method of Chouakria [6], see also ref. [4]). These are described briefly in Section 4 and compared with the vertices method. In a different direction, Palumbo and Lauro [7] and Lauro and Palumbo [8] use interval arithmetic ideas (of Moore [9]) to calculate the variance–covariance matrix based on interval means and distances. Gioia and Lauro [10] and Lauro and Gioia [11] propose an extension of classical principal components to intervals based on interval algebra properties and main results on interval eigenvalues and interval eigenvectors obtained by Deif [12] and Rhon [13]. Unfortunately, interval arithmetic ideas do not work well for principal component analysis unless the intervals are short. There is also a series of papers [14–19] which consider principal component analysis of interval fuzzy data. However, while fuzzy data can be viewed as a special case of interval data, they are in general a different domain from symbolic data; see ref. [3] for examples showing the distinctions between these two types of data.

In Section 5, we analyze a set of $m = 27$ faces dataset, with $p = 6$ variables, from Leroy *et al*. [20] investigating facial characteristics for detection purposes in a surveillance study. Facial recognition has taken on an added urgency in the last decade or so. The recent extensive review by Zhao *et al*. [21] highlights the relative paucity of statistical methodology to add to the largely computer-based methods and draws special attention to the need for techniques when databases are large. In this sense, our analysis contributes to the knowledge base for this field in that it provides a new exploratory method to aid in the process of detecting which variables are important. More importantly, however, is the wider applicability of the methodology to many fields (including the image processing field) when faced with interval-valued databases, in general. We also demonstrate, through this dataset, how attempts to analyze interval-valued data with classically valued surrogates lose information contained in the data; that is, the symbolic analysis gives more informative results than does a classical analysis, lending more importance to the usefulness of the symbolic approach.

Finally in Section 6, a second example illustrates how the method can be applied even when the number of parameters exceeds the number of observations, that is, when $m < p$, as well as how uncertainty data can be accommodated.

2. THE VERTICES PRINCIPAL COMPONENTS METHOD

2.1. Data Matrix

Suppose the data consist of m observations $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})$ where $\xi_{ij} = [a_{ij}, b_{ij}]$ with $a_{ij} \leq b_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, p$, are realizations of the random variable $\boldsymbol{X} = (X_1, \ldots, X_p)$. An interval $[a_{ij}, b_{ij}]$ is defined to be *trivial* if it reduces to a single value $a_{ij} = b_{ij}$. Notice that $\boldsymbol{\xi}_i$ is a classical observation if ξ_{ij} for all $j = 1, \ldots, p$, are trivial intervals.

Let the number of nontrivial intervals in $\boldsymbol{\xi}_i$ be q_i. Then, the number of vertices associated with the observation $\boldsymbol{\xi}_i$ in $\Re^p$ is

$$n_i = 2^{q_i}. \tag{1}$$

Thus, a classical observation which equates to a point in $\Re^p$ has $q_i = 0$ and hence has $2^0 = 1$ vertex in $\Re^p$; a line

segment has $q_i = 1$ and so has 2 vertices, a rectangle has $q_i = 2$ with $2^2 = 4$ vertices in $\Re^p$, and so on. Figure 1 displays the hypercube describing each of seven interval-valued observations measured on $p = 3$ random variables along with the corresponding clusters of vertices. We refer to all observations as being hypercubes H in $\Re^p$. The total number of vertices for the dataset $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$ is

$$n = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} 2^{q_i}. \tag{2}$$

We construct the $n_i \times p$ data matrix $\boldsymbol{X}_{\xi_i}$ with elements (x_{kj}^i), $k = 1, \ldots, n_i$, $j = 1, \ldots, p$, where $\boldsymbol{x}_k^i = (x_{k1}^i, \ldots, x_{kp}^i)$ is the point value of the vertex k, $k = 1, \ldots, n_i$, associated with the hypercube H_i representing the observation $\boldsymbol{\xi}_i$, $i = 1, \ldots, m$. Then the data matrix whose elements represent the vertices of the complete dataset is the $n \times p$ matrix

$$\boldsymbol{X} = (\boldsymbol{X}_{\xi_1}, \ldots, \boldsymbol{X}_{\xi_m})' = ((x_{kj}^1), \ldots, (x_{kj}^m))'. \tag{3}$$

2.2. Weights

As for classical analyses, there are many possible weighting schemes, generally dictated by the nature of the application at hand. We present three main symbolic weighting schemes, where without loss of generality, we assume observations have been normalized. First, let us denote the weight of observation $\boldsymbol{\xi}_i$ by w_i. A symbolic observation $\boldsymbol{\xi}_i$ has n_i vertices each of which can have a weight factor; let the weight of the vertex k (of $\boldsymbol{\xi}_i$) be w_k^i, $k = 1, \ldots, n_i$, $i = 1, \ldots, m$. Further, it follows that we require

$$w_i = \sum_{k=1}^{n_i} w_k^i, \qquad \sum_{i=1}^{m} w_i = 1. \tag{4}$$

A frequent choice of weight for w_i gives equal weight to all observations, that is,

$$w_i = 1/m, \quad i = 1, \ldots, m. \tag{5}$$

This choice for w_i gives equal weight to observations even when they have different internal variations. For example, for $p = 1$, the observations $\xi_1 = [59, 61]$ and $\xi_2 = [57, 63]$ would be equally weighted under Eq. (5).

One weighting scheme which gives importance to differing internal variations of hypercubes is given by

$$w_i = V_i \Big/ \sum_{i=1}^{m} V_i, \tag{6}$$

where V_i is the volume of the hypercube H_i associated with $\boldsymbol{\xi}_i$ given by

$$V_i = \prod_{a_{ij} \neq b_{ij}} (b_{ij} - a_{ij}). \tag{7}$$

Note that 'volume' is a generic nondimensional term, and could be a 'surface' (or 'length') in 2 (or 1) dimensions; it is simply a measure of information contained in the observation H_i. Under this scheme, observations that form larger hypercubes (and so have larger internal variability) receive larger weights. An observation that is a single point receives a weight of zero. For example, this weighting scheme might be useful when hypercubes emerge from aggregation of very large datasets, with larger hypercubes representing an aggregation of a larger number of individual observations, or more information, than smaller hypercubes. In this sense, a hypercube that is a single point in $\Re^p$ is but one distinct observation, and so its zero weight under this scheme is akin to the notion that the probability of a point is zero. Along similar but different lines, the weights w_i can be proportional to the number of observations aggregated to produce each $\boldsymbol{\xi}_i$. Another notion of 'volume' is the linear description potential equal to the sum of the hypercube edges of $\boldsymbol{\xi}_i$ [5,22].

A third scheme is one for which the weights are inversely proportional to volume, viz.,

$$w_i = \frac{1 - V_i \Big/ \sum_{i=1}^{m} V_i}{\sum_{i=1}^{m} [1 - V_i \Big/ \sum_{i=1}^{m} V_i]}. \tag{8}$$

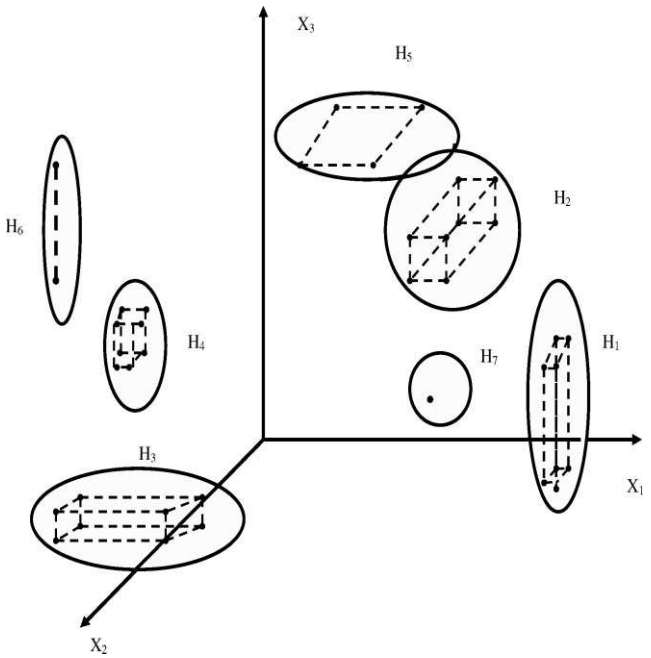

Fig. 1 Types of hypercubes: clouds of vertices.

In this case, observations with large volumes receive lower weight. This type of weighting scheme might be more appropriate for observations if the intervals are measures of imprecision ($x \pm \delta$), with lower weights for more uncertainty (i.e., larger δ) expressed through the observation's interval range.

Consider now the weights for each vertex k. When the weights for the n_i vertices of the observation $\boldsymbol{\xi}_i$ (or, hypercube H_i) are assumed to be equal,

$$w_k^i = w_i/n_i, \quad k = 1, \ldots, n_i, \quad i = 1, \ldots, m. \tag{9}$$

For example, for $\boldsymbol{\xi}_1 = ([3,\ 5],\ [10,\ 16],\ [7,\ 9])$, $\boldsymbol{\xi}_2 = ([3,\ 5],\ [13,\ 13],\ [8,\ 8])$, these become $w_i = 1/2$ with $w_k^1 = 1/16,\ w_k^2 = 1/4,\ k = 1, \ldots, n_i$, since from Eq. (1) $n_1 = 8$ and $n_2 = 2$.

When nothing is known about the internal distribution across an interval, these weights could be determined with regard to the means located at the interval midpoints $x_{ij}^c = (a_{ij} + b_{ij})/2$, which in effect is assuming a uniform distribution within the intervals. More generally for any distribution, rather then the midpoints, some suitably defined reference point x_{ij}^0 with $a_{ij} < x_{ij}^0 < b_{ij}$ can be used. For example, x_{ij}^0 can be the mode, it can be the observed mean across $[a_{ij}, b_{ij}]$ for the distribution underlying the corresponding X_j, or so on. We then set weights w_{ij}^a and w_{ij}^b on the endpoints a_{ij} and b_{ij}, respectively, such that

$$w_{ij}^a + w_{ij}^b = 1 \quad \text{and} \quad w_{ij}^a a_{ij} + w_{ij}^b b_{ij} = x_{ij}^0. \tag{10}$$

Then, the weights w_k^i for the vertex k of $\boldsymbol{\xi}_i$ can be given by

$$w_k^i = w_i \left[\prod_{j=1}^{q_i} w(x_{kj}^i) \right], \tag{11}$$

where the weight associated with the jth component of the k vertex is

$$w(x_{kj}^i) = w_{ij}^t, \ \text{when } x_{kj} = t_{ij}, \ t = a, b. \tag{12}$$

To illustrate, consider a $p = 2$ observation $\boldsymbol{\xi}_i = ([a_{i1}, b_{i1}],\ [a_{i2}, b_{i2}])$ which forms a rectangle hypercube H_i with $n_i = 4$ vertices. Then for the $k = 1, \ldots, 4$ vertices, we have

$$w_1^i = w_i w_{i1}^a w_{i2}^a, \quad w_2^i = w_i w_{i1}^a w_{i2}^b,$$
$$w_3^i = w_i w_{i1}^b w_{i2}^a, \quad w_4^i = w_i w_{i1}^b w_{i2}^b.$$

Then, after applying Eq. (10) for each of $j = 1, 2$, we have

$$\sum_{k=1}^{4} w_k^i = w_i\{w_{i1}^a(w_{i2}^a + w_{i2}^b) + w_{i1}^b(w_{i2}^a + w_{i2}^b)\} = w_i.$$

For the case that x_{ij}^0 is the interval midpoint x_{ij}^c, the weights $w_{ij}^a = w_{ij}^b = 1/2$, and so the particular weights of Eq. (9) pertain.

It follows that the weight matrix $\boldsymbol{D}$ associated with the observation vertices matrix $\boldsymbol{X}$ is the $n \times n$ diagonal matrix

$$\boldsymbol{D} = \text{diag}(w_1^1, \ldots, w_{n_1}^1, \ldots, w_1^m, \ldots, w_{n_m}^m). \tag{13}$$

2.2.1. *Classical data*

When all the observations are classical data with $a_{ij} = b_{ij}$ for all $i = 1, \ldots, m,\ j = 1, \ldots, p$, it follows that the number of nontrivial intervals $q_i = 0$ and hence $n_i = 1$ for all $i = 1, \ldots, m$. Hence, the vertex weights $w_k^i \equiv w_i$. All the results herein carry through as a special case.

2.3. Variance–Covariance Matrix

Principal component analysis includes finding the eigenvalues and eigenvectors of the variance–covariance matrix of the data. Let us define the variance–covariance matrix associated with the vertices by $\boldsymbol{V} = (v_{j_1, j_2}),\ j_1, j_2 = 1, \ldots, p$; then,

$$\boldsymbol{V} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{X}, \tag{14}$$

where $\boldsymbol{X}$ and $\boldsymbol{D}$ are as defined in Eqs. (3) and (13), respectively. Recalling that the structure of the matrix $\boldsymbol{X}$ is that it represents the n vertex points in the complete symbolic dataset and can be viewed as n classical observations, we can obtain the weighted sample means as

$$\overline{X}_j = \sum_{i=1}^{m} \sum_{k=1}^{n_i} w_k^i x_{kj}^i = \sum_{i=1}^{m} (\alpha_{ij}^a a_{ij} + \alpha_{ij}^b b_{ij}), \tag{15}$$

where α_{ij}^a and α_{ij}^b are the weights for the observation $\boldsymbol{\xi}_i$ when the value of x_{kj}^i is a_{ij} and b_{ij}, respectively. Therefore, for $t = a, b$,

$$\alpha_{ij}^t = \sum_{k=1}^{n_i} w_k^i = w_{ij}^t w_i \ \text{whenever } x_{kj}^i = t_{ij}. \tag{16}$$

It follows from Eq. (12) that $\alpha_{ij}^a + \alpha_{ij}^b = w_i$.

Then, the variance v_{jj} of X_j can be written as

$$v_{jj} = \sum_{i=1}^{m} \sum_{k=1}^{n_i} w_k^i (x_{kj}^i - \overline{X}_j)^2; \tag{17}$$

hence,

$$v_{jj} = \sum_{i=1}^{m} [\alpha_{ij}^a (a_{ij} - \overline{X}_j)^2 + \alpha_{ij}^b (b_{ij} - \overline{X}_j)^2]. \tag{18}$$

Likewise, the covariance $v_{j_1 j_2}$ between X_{j_1} and X_{j_2} can be written as

$$v_{j_1 j_2} = \sum_{i=1}^{m} \sum_{k=1}^{n_i} w_k^i (x_{kj_1}^i - \overline{X}_{j_1})(x_{kj_2}^i - \overline{X}_{j_2}). \quad (19)$$

We can show that

$$\begin{aligned} v_{j_1 j_2} &= \sum_{i=1}^{m} w_i \\ &(w_{ij_1}^a w_{ij_2}^a a_{ij_1} a_{ij_2} + w_{ij_1}^a w_{ij_2}^b a_{ij_1} b_{ij_2} + w_{ij_1}^b w_{ij_2}^a b_{ij_1} a_{ij_2} \\ &\quad + w_{ij_1}^b w_{ij_2}^b b_{ij_1} b_{ij_2}) \\ &= \sum_{i=1}^{m} w_i x_{ij_1}^0 x_{ij_2}^0 \end{aligned} \quad (20)$$

from Eq. (10). Hence, the variance–covariance matrix $\boldsymbol{V}$ based on the vertices is calculated.

2.3.1. Complexity

Since each observation is represented by its n_i vertices, it would seem that when calculating the variance–covariance matrix $\boldsymbol{V}$, the order of complexity is $O(m2^p)$ when there are no trivial intervals. If p is large, this can be considerable. However, this complexity can be reduced to $O(m)$ by consideration of the relevant variance–covariance matrix $\mathbf{V^c}$ obtained from the reference points x_{ij}^0 introduced in Section 2.2.

We can show that these reference points (e.g., midpoints) have weighted mean

$$\overline{X}_j^c = \sum_{i=1}^{m} w_i x_{ij}^0 = \sum_{i=1}^{m} (\alpha_{ij}^a a_{ij} + \alpha_{ij}^b b_{ij}),$$

that is, $\overline{X}_j^c = \overline{X}_j$ from Eq. (15). We can also show that $\mathbf{V^c}$ has elements $v_{j_1 j_2}^c$ given by

$$v_{j_1 j_2}^c = \sum_{i=1}^{m} w_i (w_{ij_1}^a a_{ij_1} + w_{ij_1}^b b_{ij_1})(w_{ij_2}^a a_{ij_2} + w_{ij_2}^b b_{ij_2}). \quad (21)$$

Now, when $j_1 = j_2 = j$, from Eqs. (20) and (21), we can show that

$$v_{jj} = v_{jj}^c + e_{jj}, \quad j = 1, \ldots, p, \quad (22)$$

where

$$e_{jj} = \sum_{i=1}^{m} w_i w_{ij}^a w_{ij}^b (b_{ij} - a_{ij})^2. \quad (23)$$

Likewise, from Eqs. (20) and (21) when $j_1 \neq j_2$, we can show that $v_{j_1 j_2} = v_{j_1 j_2}^c$.

Hence, the vertices variance–covariance matrix $\boldsymbol{V}$ and the variance–covariance matrix $\boldsymbol{V}^c$ satisfy the relationship

$$\boldsymbol{V} = \boldsymbol{V}^c + \boldsymbol{E}, \quad (24)$$

where $\boldsymbol{E}$ is a $p \times p$ diagonal matrix with diagonal elements e_{jj} given by Eq. (23). Note that $\boldsymbol{V}^c$ represents the between observations variation and $\boldsymbol{E}$ describes a within observations or internal variations of the data.

The relationship given in Eq. (24) allows for the calculation of the vertices variance–covariance matrix $\boldsymbol{V}$ by calculating the variance–covariance matrix $\boldsymbol{V}^c$ and the difference matrix $\boldsymbol{E}$, with complexity $O(m)$, instead of the complexity $O(m2^p)$ that pertains when calculating $\boldsymbol{V}$ directly through the vertices as in Eq. (14). Therefore, the degree of complexity for the vertices method is reduced to the order $O(m)$, the same as for a classical principal component analysis.

2.3.2. Classical data

When the data are all classical observations, we have from Eq. (23), $e_{jj} = 0$, for all j. In this case, the two methods are equivalent throughout and the classical principal component analysis becomes a special case.

2.3.3. Centers method

Cazes *et al.* [4] also provided a brief summary of a so-called centers method. In this case, the reference points x_{ij}^0 are the interval midpoints, that is, $x_{ij}^0 = x_{ij}^c$. Therefore, the variance–covariance matrix is just $\mathbf{V^c}$, and so ignores internal variations contained in the data (expressed through $\boldsymbol{E}$). See ref. [6] for details of this method.

2.4. Vertices Principal Component Analysis

The data matrix $\boldsymbol{X}$ is a data matrix of n classical point observations on the random variables $(X_1, \ldots, X_p)$, with its associated weighted variance–covariance matrix $\boldsymbol{V}$ of Eq. (14). Therefore, we can perform a classical principal component analysis on this $\boldsymbol{X}$. A detailed description of how to conduct such an analysis can be found from any of the numerous texts on multivariate analysis; see, e.g., refs. [1,23] for an applied presentation, and ref. [24] for a theoretical approach.

From Cazes *et al.* [4], the νth symbolic vertices principal component for the symbolic observation $\boldsymbol{\xi}_i$ represented by the n_i vertices in $\boldsymbol{X}_{\xi_i}$ is

$$Y_{i\nu}^* = [y_{i\nu}^a, y_{i\nu}^b], \quad \nu = 1, \ldots, s \leq p, \quad (25)$$

where

$$y_{iv}^a = \min_{k\in L_i}\{y_{vk}^i\}, \quad y_{iv}^b = \max_{k\in L_i}\{y_{vk}^i\}, \tag{26}$$

where $L_i = \{1, \ldots, n_i\}$ is the set of rows in $\boldsymbol{X}_{\xi_i}$ which describe the vertices of the symbolic hypercube H_i and hence the observation $\boldsymbol{\xi}_i$, and where y_{vk}^i is the value of the vth principal component for the row k in L_i.

This result can be verified as follows. Defining $J^+ = \{j|e_{vj} > 0\}$ and $J^- = \{j|e_{vj} < 0\}$, we can show that

$$y_{iv}^a = \sum_{j\in J^+}^{p} e_{vj}(a_{ij} - \overline{X}_j) + \sum_{j\in J^-}^{p} e_{vj}(b_{ij} - \overline{X}_j), \tag{27}$$

$$y_{iv}^b = \sum_{j\in J^-}^{p} e_{vj}(a_{ij} - \overline{X}_j) + \sum_{j\in J^+}^{p} e_{vj}(b_{ij} - \overline{X}_j), \tag{28}$$

where $\boldsymbol{e}_v = (e_{v1}, \ldots, e_{vp})$, $v = 1, \ldots, p$, is the vth eigenvector associated with the vth eigenvalue (λ_v) of $\boldsymbol{V}$. Next, take any point $\tilde{\boldsymbol{x}}_i$ with $\tilde{x}_{ij} \in [a_{ij}, b_{ij}]$. Then, the vth principal component associated with this $\tilde{\boldsymbol{x}}_i$ is

$$\tilde{\mathrm{PC}}v = \sum_{j=1}^{p} e_{vj}(\tilde{x}_{ij} - \overline{X}_j).$$

It follows that

$$\sum_{j=1}^{p} e_{vj}(\tilde{x}_{ij} - \overline{X}_j) \geq \sum_{j\in J^+}^{p} e_{vj}(a_{ij} - \overline{X}_j) + \sum_{j\in J^-}^{p} e_{vj}(b_{ij} - \overline{X}_j) \tag{29}$$

and

$$\sum_{j=1}^{p} e_{vj}(\tilde{x}_{ij} - \overline{X}_j) \leq \sum_{j\in J^-}^{p} e_{vj}(a_{ij} - \overline{X}_j) + \sum_{j\in J^+}^{p} e_{vj}(b_{ij} - \overline{X}_j). \tag{30}$$

However, by definition Eq. (26) and from Eqs. (27 and (28), the right-hand side of Eqs. (29) and (30) are, respectively,

$$\min_{k\in L_i}\{y_{vk}^i\} = y_{iv}^a, \qquad \max_{k\in L_i}\{y_{vk}^i\} = y_{iv}^b.$$

Hence, for all $v = 1, \ldots, p$, $\tilde{\mathrm{PC}}v \in [y_{iv}^a, y_{iv}^b]$; and so Y_{iv}^* as in Eqs. (25) and (26) holds for all $x_{ij} \in [a_{ij}, b_{ij}]$.

A graphical representation of a set of $s = 2$ principal components obtained from the projection of the hypercube H_i with $n_i = 6$ vertices onto the PC1 and PC2 plane is displayed in Fig. 2. The rectangle formed by the two interval-valued principal components constitutes a maximal envelope of the projection points from H_i. Thus, every point in the hypercube H_i when projected to the plane lies inside this envelope. However, depending on the actual value of H_i, there can be some (exterior) points within the envelope that may not be projections of points in H_i. In this sense, the envelope overestimates the principal component hypercube. This can be improved by looking at the quality of each vertex, as introduced in Section 3.

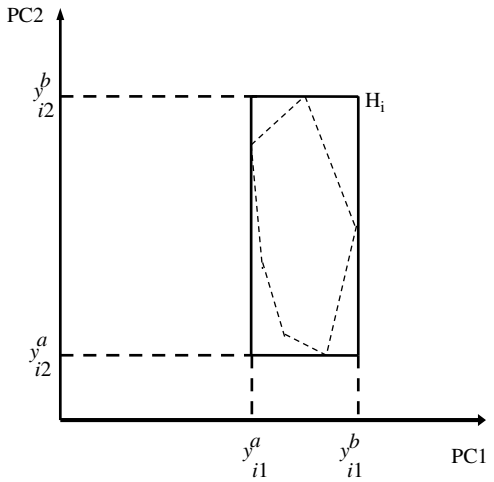

Fig. 2 Projection hypercube H_i to principal component ($v = 1, 2$) axes.

As for classical analyses, the total variance is $\sigma_n^2 = \sum_{i=1}^{p} \lambda_i$, $\mathrm{Var}(\mathrm{PC}v) = \lambda_v$ and the proportion of the total variance explained by PCv is $\lambda_v / \sum_{v=1}^{p} \lambda_v$. Also, we can obtain a correlation measure between the vth principal component PCv and the random variable X_j as

$$C_{jv} = \mathrm{Cor}(X_j, \mathrm{PC}v) = e_{vj}\sqrt{\lambda_v/\sigma_j^2}, \tag{31}$$

where σ_j^2 is the variance of X_j. Note that when the variance–covariance matrix is standardized, these σ_j^2 reduce to $\sigma_j^2 = 1$.

3. INTERPRETATION AND VISUALIZATION

In classical principal component analysis, two different quantities are usually calculated to help in the visualization and interpretation of the projections of the principal component values for each observation onto the principal component axes. One is the cosine of each (classical) observation X_i onto the vth principal component axis, viz.,

$$\cos(X_i, \mathrm{PC}v) = w_i y_{iv}^2/[d(X_i, G)]^2,$$

where $d(X_i, G)$ is the Euclidean distance between the observation X_i and G is the centroid of all data X_i, $i = 1, \ldots, n$, values. Large values of $\cos(X_i, \mathrm{PC}v)$ mean that

the position of X_i is near to its projected value on the PCν axis and hence we are confident about the position of this X_i observation's role in the interpretation of the principal component analysis results; low values of $\cos(X_i, \text{PC}\nu)$ suggest care is necessary when interpreting results relative to that X_i and PCν. A second quantity useful for interpretation purposes in a classical analysis is the contribution of each observation X_i to the inertia, viz.,

$$\text{Ctr}(X_i, \text{PC}\nu) = w_i y_{i\nu}^2/\lambda_\nu.$$

For our symbolic principal component analyses, instead of a single point observation X_i, we have the hypercube H_i. We extend these classical quantities to hypercubes as follows. The relative contribution to a given principal component PCν by an observation $\boldsymbol{\xi}_i$ represented here by its observed hypercube H_i can be measured by

$$C_{i\nu}^1 = \text{Ctr}(H_i, \text{PC}\nu) = w_i \sum_{k=1}^{n_i} \frac{w_k^i (y_{\nu k}^i)^2}{[d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2}, \tag{32}$$

where $y_{\nu k}^i$ is the νth principal component for the vertex k of H_i (see Eq. (26)), w_k^i is the weight of that vertex (see Section 2.2), and where $d(\boldsymbol{x}_k^i, \boldsymbol{G})$ is the Euclidean distance between the vertex $\boldsymbol{x}_k^i$ identified in the row k of $\boldsymbol{X}_{\xi_i}$ and $\boldsymbol{G}$ defined as the centroid of all n rows of $\boldsymbol{X}$. An alternative measure is the contribution

$$C_{i\nu}^2 = \text{Ctr}(H_i, \text{PC}\nu) = \frac{\sum_{k=1}^{n_i} w_k^i (y_{\nu k}^i)^2}{\sum_{k=1}^{n_i} w_k^i [d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2}. \tag{33}$$

The first function $C_{i\nu}^1$ identifies the average squared cosines of the angles between these vertices and the axis of the νth principal component. The second function $C_{i\nu}^2$ identifies the ratio between the contribution of all the vertices of H_i to the variance λ_ν of the νth principal component and their contribution to the total inertia (or total variance). Also, since for all positive real numbers a, b, c, d, the relationship $(a + c)/(b + d) \leq [a/b + c/d)$ holds, then it follows that $C_{i\nu}^2 \leq C_{i\nu}^1$.

The absolute contribution of a single observation through the vertices of H_i to the variance of PC$\nu = \lambda_\nu$ is measured by the inertia

$$I_{i\nu} = \text{Inertia}(H_i, \text{PC}\nu) = \left[\sum_{k=1}^{n_i} w_k^i (y_{\nu k}^i)^2 \right] / \lambda_\nu, \tag{34}$$

and the contribution of this observation to the total variance is

$$I_i = \text{Inertia}(H_i) = \left\{ \sum_{k=1}^{n_i} w_k^i [d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2 \right\} / I_\text{T}, \tag{35}$$

where $I_\text{T} = \sum_{\nu=1}^{p} \lambda_\nu$ is the total variance of all the vertices in $\Re^p$. It is easily verified that

$$\sum_{i=1}^{m} I_{i\nu} = \lambda_\nu, \quad \sum_{i=1}^{m} I_i = I_\text{T}. \tag{36}$$

An alternative visual aid in interpreting the results is that only those vertices whose contribution to the principal component PCν exceed some prespecified value α, be used in Eq. (25). That is, we set

$$Y_{i\nu}^*(\alpha) = [y_{i\nu}^a(\alpha), y_{i\nu}^b(\alpha)],$$

where

$$\begin{aligned} y_{i\nu}^a(\alpha) &= \min_{k \in L_i} \{ y_{\nu k}^i | \text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu) \geq \alpha \}, \\ y_{i\nu}^b(\alpha) &= \max_{k \in L_i} \{ y_{\nu k}^i | \text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu) \geq \alpha \}, \end{aligned} \tag{37}$$

where

$$\text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu) = \frac{(y_{\nu k}^i)^2}{[d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2} \tag{38}$$

is the contribution of a single vertex $\boldsymbol{x}_k^i$ to the νth principal component.

When $\alpha = 0$, the principal component interval obtained from Eq. (26) has an underlying assumption that all n vertices are equally important in determining that interval regardless of the respective contributions of individual vertices calculated from Eq. (37) or (38). Thus, to take an extreme case, one vertex $k = k'$ may have a value of $y_{\nu k'}^i = 1.0$ (say) while all the other vertices $k \neq k'$ in L_i may take values in the range 10.0, ... ,11.0 (say) for a given value of ν. Direct use of Eq. (26) gives PC$\nu = [1.0,\ 11.0]$. Suppose however the relative contribution, from Eq. (37), for that vertex k' is 0.05, while those for the other vertices $k \neq k'$ in L_i are such that they exceed $\alpha = 0.6$ (say). Then, a more meaningful symbolic principal component interval in this case is PC$\nu = [10.0,\ 11.0]$. On the other hand, if the $k = k'$ vertex has a contribution of 0.65 (say), then now all vertices should be included, and so from Eq. (37), we have PC$\nu = [1.0,\ 11.0]$. That is, if a particular vertex contributes relatively little information to a specific principal component calculation, it is omitted from Eq. (26) allowing only those vertices which are meaningful to be retained. For classical data, there is only one vertex ($n = 1$) and so this argument does not arise.

An alternative to the criterion of Eq. (37) is to replace $\text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu)$ by

$$\text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu_1, \text{PC}\nu_2) = \text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu_1) + \text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu_2). \tag{39}$$

In this case, vertices that make larger contributions in either of the two principal components PCν_1 and PCν_2 are retained, rather than only those vertices that contribute to just one principal component.

To illustrate, consider the projections of the two hypercubes H_1 and H_2 onto the first and second principal component plane as shown in Fig. 3. The principal component envelope (obtained from applying Eq. (26)) is also displayed. The numerical values at each of the projected vertices are the contributions of the respective vertices to the first ($\nu = 1$) principal component, calculated from Eq. (38). For example, the five vertices of H_1, respectively, contribute 0.8, 0.05, 0.55, 0.35, 0.75, to PC1. When $\alpha = 0.2$ (say) in Eq. (37), the vertex contributing 0.05, is omitted, with the resulting principal component envelope being that shown in Fig. 4. The observation represented by the hypercube H_2 has six vertices (see Fig. 3) including a vertex whose contribution is 0.01. Application of Eq. (37) when $\alpha = 0.2$ results in the two vertices whose contributions are 0.01 and 0.15 being dropped. However, the resulting principal component envelope in this case still includes the vertex with contribution 0.01.

When for a given $\nu = \nu_1$ (say) all $\mathrm{Ctr}(\boldsymbol{x}_k^i, \mathrm{PC}\nu_1) < \alpha$, then to keep track of the position of the hypercube H_i on the principal component plane (ν_1, ν_2) say, we project the center of the hypercube onto that axis at

$$\overline{y}_{i\nu_1} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{\nu_1 k}^i. \tag{40}$$

In this case, there is no variability on that principal component ν_1; whereas if there is variability for the other principal component ν_2, there is a line segment on its (ν_2) plane.

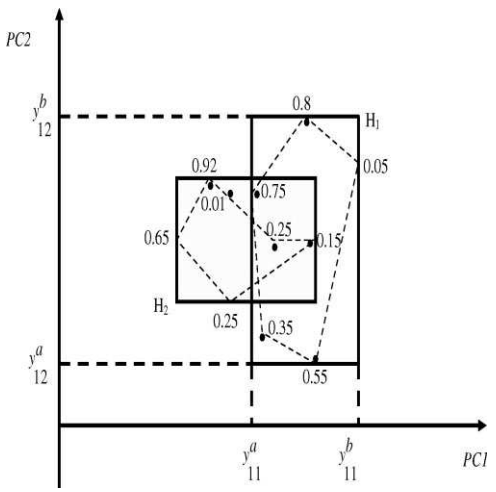

Fig. 3 Principal component envelope, $\alpha = 0$, based on relative contributions of vertices to PCν, $\nu = 1, 2$.

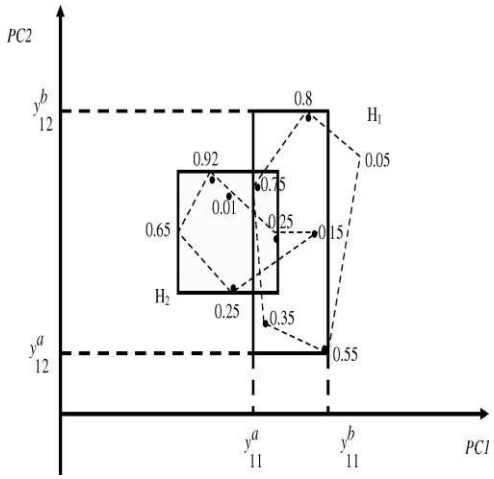

Fig. 4 Principal component envelope, $\alpha = 0.2$, based on relative contributions of vertices to PCν, $\nu = 1, 2$.

4. RANGE BASED METHODS

In an attempt to improve the factorial visualization of the vertices method, Lauro and Palumbo [5] describe three variants of the Cazes *et al.* [4] centers method. It is assumed all intervals are nontrivial. First they introduced a so-called Symbolic Object—Principal Component Analysis (SO-PCA) method in which they performed a principal components analysis on the interval midpoints. That is, they used the variance–covariance matrix $\boldsymbol{V}^c$ of Section 2.3 for the particular case that the reference points x_{ij}^0 are the midpoints x_{ij}^c for the observed intervals. Principal component envelopes are then constructed by substituting the resulting eigenvector values (from $\boldsymbol{V}^c$) into Eqs. (27) and (28) and by replacing $\overline{X}_j$ by $\overline{X}_j^c$. As for the centers method, in this SO-PCA method, principal component axes are totally defined by the midpoints of the hypercubes. The midpoints are standardized by a variance calculation through a boolean matrix based on the $m2^p$ vertices. Therefore, this SO-PCA method has a computational complexity of $O(m2^p)$ rather than the complexity of only $O(m)$ for the pure centers method (see Section 2.3). Lauro and Palumbo recognized this complexity as problematic, calling it 'the curse of dimensionality'. Note that because of Eq. (24), the vertices method does not have this 'curse'.

Lauro and Palumbo [5] further proposed two additional methods based on the range, called the Range-Transformation-PCA (RT-PCA) method and the Mixed-PCA method (which combines the SO-PCA and RT-PCA methods). The RT method performs a classical principal component analysis on the maximum vertices after being centered on the minimum vertices. The Mixed-PCA method performs a principal component analysis on the recoded midpoints obtained by projecting the midpoints of the observations onto the factorial axes found from the RT-PCA

method. Another method using the range variable is that of Giordani and Kiers [17] in their analysis of fuzzy data. This approach replaces each interval by two random variables, the midpoint and range values. Their approach is considered in Section 5.3 when looking at classical surrogates of the interval-valued data. The underlying problem identified there for ranges applies also to the Lauro and Palumbo [5] method.

5. FACE RECOGNITION APPLICATION

5.1. The Data

The problem of automatic face recognition has gained added impetus recently especially in the context of security such as in access to buildings and the like, and in the context of monitoring and continued surveillance questions. Mechanisms for identifying human facial patterns started receiving attention with the Fischler and Eschlager [25] study of matching pictorial structures, followed by Baron [26], among others. Following a brief review by Samal and Iyengar [27], in an excellent and extensive review, Chellappa *et al*. [28] looks at face recognition in the law enforcement and commercial sectors as well as the psychophysics community. The last ten years has witnessed considerable activity on this vexing issue. Zhao *et al*. [21] provides an in-depth review of the recent literature. Much of this work falls under the broad rubric of image analysis; while some deal with computer architectural graph matching methods. There are a few studies involving direct statistical methods, such as principal component analysis of eigenfaces used by Turk and Pentland [29], Craw and Cameron [30], and Moon and Phillips [31], discriminant analysis by Eternad and Chellappa [32], probabilistic eigenfaces developed by Moghaddam and Pentland [33], and nearest line features considered by Li and Chellappa [34] and Li and Lu [35]. Studies such as those by Kass *et al*. [36], Turk [37], Craw *et al*. [38], and Staib and Duncan [39] helped identify those facial features that should be included in any discrimination research. Zhao *et al*. [21] conclude that while progress has been valuable, much more remains to be done especially when databases are large.

Our analysis focuses on a dataset from an investigation by Leroy *et al*. [20] which uses face recognition features identified from these earlier studies. The process of face recognition entails first describing the faces, then classifying and lastly identifying them. One technique for describing faces consists of taking a number of measurements, which identify principal facial features (width of eyes, nose, etc.). The classification stage is achieved through a principal component analysis to identify groupings of faces with the associated interpretations providing input as to the identification of distinguishing features. Our methodology provides a new exploratory technique when the data are intervals instead of the points of classical data.

The dataset consists of measurements of six random variables designed to identify each face; specifically, the length spanned by the eyes X_1 (the distance AD in Fig. 5), the length between the eyes X_2 (the distance BC), the length from the outer right eye to the upper middle lip at the point H between the nose and mouth X_3 (AH), the corresponding length for the left eye X_4 (DH), the length from this point H to the outside of the mouth on the right side X_5 (EH) and the corresponding distance to the left side of the mouth X_6 (GH). For each face image, the localization of the salient features such as nose, mouth, and eyes is obtained by using morphological operators. In order to extract the boundary of these localized elements, a specific active contour method based on Fourier descriptors able to incorporate information about the global shape of each object is used. Finally, specific points delimiting the extracted boundaries are localized, and then a distance is measured between a specific pair of points as represented by these random variables, in Fig. 5. This distance measure is expressed as the number of pixels on an image of the face. There is a sequence of such images; so therefore the actual distances measured are interval-valued. Thus, for example, the eye-span distance X_1 for the subject FRA1 is $X_1 = [155.00, 157.00]$ over this series of images. Note that due to the different conditions of alignment, illumination, pose and occlusion, the extracted distances will vary across the different images of the same person. The study involved nine men with three sequences for each giving a total of $m = 27$ observations. The complete dataset is provided in Table 1.

Before carrying out the analysis, let us first make the following comment that pertains for aggregated data such as these faces data. As described, there are 27 interval-valued observations. Suppose each observation

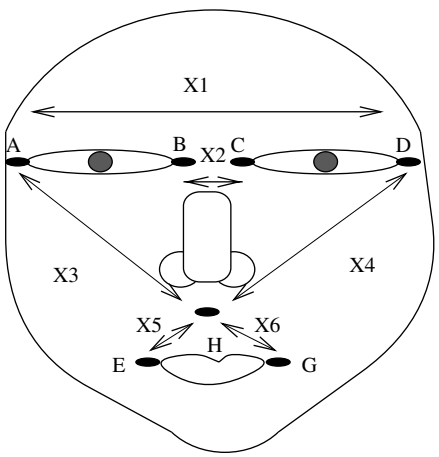

Fig. 5 Face: description of random variables.

Table 1. Faces dataset (distances AD, ... ,GH as in Fig. 5, see text).

Subject	$X_1 = \text{AD}$	$X_2 = \text{BC}$	$X_3 = \text{AH}$	$X_4 = \text{DH}$	$X_5 = \text{EH}$	$X_6 = \text{GH}$
FRA1	[155.00, 157.00]	[58.00, 61.01]	[100.45, 103.28]	[105.00, 107.30]	[61.40, 65.73]	[64.20, 67.80]
FRA2	[154.00, 160.01]	[57.00, 64.00]	[101.98, 105.55]	[104.35, 107.30]	[60.88, 63.03]	[62.94, 66.47]
FRA3	[154.01, 161.00]	[57.00, 63.00]	[99.36, 105.65]	[101.04, 109.04]	[60.95, 65.60]	[60.42, 66.40]
HUS1	[168.86, 172.84]	[58.55, 63.39]	[102.83, 106.53]	[122.38, 124.52]	[56.73, 61.07]	[60.44, 64.54]
HUS2	[169.85, 175.03]	[60.21, 64.38]	[102.94, 108.71]	[120.24, 124.52]	[56.73, 62.37]	[60.44, 66.84]
HUS3	[168.76, 175.15]	[61.40, 63.51]	[104.35, 107.45]	[120.93, 125.18]	[57.20, 61.72]	[58.14, 67.08]
INC1	[155.26, 160.45]	[53.15, 60.21]	[95.88, 98.49]	[91.68, 94.37]	[62.48, 66.22]	[58.90, 63.13]
INC2	[156.26, 161.31]	[51.09, 60.07]	[95.77, 99.36]	[91.21, 96.83]	[54.92, 64.20]	[54.41, 61.55]
INC3	[154.47, 160.31]	[55.08, 59.03]	[93.54, 98.98]	[90.43, 96.43]	[59.03, 65.86]	[55.97, 65.80]
ISA1	[164.00, 168.00]	[55.01, 60.03]	[120.28, 123.04]	[117.52, 121.02]	[54.38, 57.45]	[50.80, 53.25]
ISA2	[163.00, 170.00]	[54.04, 59.00]	[118.80, 123.04]	[116.67, 120.24]	[55.47, 58.67]	[52.43, 55.23]
ISA3	[164.01, 169.01]	[55.00, 59.01]	[117.38, 123.11]	[116.67, 122.43]	[52.80, 58.31]	[52.20, 55.47]
JPL1	[167.11, 171.19]	[61.03, 65.01]	[118.23, 121.82]	[108.30, 111.20]	[63.89, 67.88]	[57.28, 60.83]
JPL2	[169.14, 173.18]	[60.07, 65.07]	[118.85, 120.88]	[108.98, 113.17]	[62.63, 69.07]	[57.38, 61.62]
JPL3	[169.03, 170.11]	[59.01, 65.01]	[115.88, 121.38]	[110.34, 112.49]	[61.72, 68.25]	[59.46, 62.94]
KHA1	[149.34, 155.54]	[54.15, 59.14]	[111.95, 115.75]	[105.36, 111.07]	[54.20, 58.14]	[48.27, 50.61]
KHA2	[149.34, 155.32]	[52.04, 58.22]	[111.20, 113.22]	[105.36, 111.07]	[53.71, 58.14]	[49.41, 52.80]
KHA3	[150.33, 157.26]	[52.09, 60.21]	[109.04, 112.70]	[104.74, 111.07]	[55.47, 60.03]	[49.20, 53.41]
LOT1	[152.64, 157.62]	[51.35, 56.22]	[116.73, 119.67]	[114.62, 117.41]	[55.44, 59.55]	[53.01, 56.60]
LOT2	[154.64, 157.62]	[52.24, 56.32]	[117.52, 119.67]	[114.28, 117.41]	[57.63, 60.61]	[54.41, 57.98]
LOT3	[154.83, 157.81]	[50.36, 55.23]	[117.59, 119.75]	[114.04, 116.83]	[56.64, 61.07]	[55.23, 57.80]
PHI1	[163.08, 167.07]	[66.03, 68.07]	[115.26, 119.60]	[116.10, 121.02]	[60.96, 65.30]	[57.01, 59.82]
PHI2	[164.00, 168.03]	[65.03, 68.12]	[114.55, 119.60]	[115.26, 120.97]	[60.96, 67.27]	[55.32, 61.52]
PHI3	[161.01, 167.00]	[64.07, 69.01]	[116.67, 118.79]	[114.59, 118.83]	[61.52, 68.68]	[56.57, 60.11]
ROM1	[167.15, 171.24]	[64.07, 68.07]	[123.75, 126.59]	[122.92, 126.37]	[51.22, 54.64]	[49.65, 53.71]
ROM2	[168.15, 172.14]	[63.13, 68.07]	[122.33, 127.29]	[124.08, 127.14]	[50.22, 57.14]	[49.93, 56.94]
ROM3	[167.11, 171.19]	[63.13, 68.03]	[121.62, 126.57]	[122.58, 127.78]	[49.41, 57.28]	[50.99, 60.46]

drew from a sequence of 1000 images. This gives a total of 27 000 classical point observations in $\Re^6$. An underlying assumption of the standard classical analysis is that all 27 000 observations are independent. However, this is not what we have here. The data values for each face form a set of 1000 dependent observations. Therefore, if we use each image as the statistical unit by performing a classical analysis, we lose the information on dependency contained in the 27 000 observations. The resulting principal component analysis will look for axes which maximize the variability across all 27 000 images regardless of whether some images belong to the same sequence. In contrast, by using the interval-valued observations obtained from each sequence, the vertices method will extract principal component axes which maximize the variability of each interval (i.e., maximizes the internal variability) and hence retains the information on dependency between the 1000 images of each sequence.

5.2. Vertices Principal Components Analysis

We first apply the vertices principal component method to these data. Observations and their vertices were given equal weights (Eqs. (8) and (12)). Values of the first three vertices principal components obtained through the application of Eq. (26) for each observation are displayed in Table 2. The plots of these along the first principal component (PC1) and second principal component (PC2) axes are shown in Fig. 6. An immediate observation is the proximity of the three sequences for the three faces for each individual thus validating their within-subject coherence. Furthermore, we can distinguish four, or possibly five, classes of faces.

By restricting the calculation of the principal components to those vertices which have a contribution α or more, that is, using Eq. (37), we can obtain a clearer picture of the class groupings. The relative contribution $\text{Ctr}(\boldsymbol{x}_k^i, \text{PC}\nu)$, $\nu = 1, 2$, $k = 1, \ldots, n_i$, are calculated from Eq. (38) for each hypercube H_i, $i = 1, \ldots, 27$. Take the face INC2 ($i = 8$) hypercube. For $\nu = 1$, all the vertices have a relative contribution $\text{Ctr}(\boldsymbol{x}_k^i, \text{PC1}) > 0.2$. Therefore, all (of the $2^6 = 64$ total) vertices enter into the application of Eq. (37) to give us $\text{PC1}(\alpha = 0.2) = [-3.662, -1.354]$. However, for $\nu = 2$, only 8 of the 64 vertices satisfy the relation $\text{Ctr}(\boldsymbol{x}_k^i, \text{PC2}) > 0.2$. Those relative contributions which satisfy this relation for the vertices of the face INC2, are given in Table 3. Therefore, only these vertices are considered in the application of Eq. (37). Hence, we

Table 2. Vertices principal components, $\nu = 1, 2, 3$: faces.

Subject	PC1	PC2	PC3
FRA1	[−2.66, −1.61]	[0.27, 1.57]	[−0.29, 1.00]
FRA2	[−2.49, −1.03]	[−0.11, 1.61]	[−0.25, 1.01]
FRA3	[−2.99, −0.81]	[−0.40, 1.88]	[−0.88, 1.20]
HUS1	[−0.24, 1.10]	[0.39, 2.05]	[0.64, 2.13]
HUS2	[−0.40, 1.41]	[0.56, 2.65]	[0.29, 2.32]
HUS3	[−0.24, 1.42]	[0.43, 2.52]	[0.27, 2.17]
INC1	[−3.77, −2.29]	[−0.67, 1.23]	[−0.80, 0.69]
INC2	[−3.66, −1.35]	[−2.05, 0.92]	[−0.88, 1.83]
INC3	[−4.02, −1.86]	[−1.20, 1.41]	[−1.01, 1.50]
ISA1	[0.80, 2.00]	[−1.83, −0.46]	[−0.58, 0.58]
ISA2	[0.37, 1.86]	[−1.71, −0.08]	[−0.64, 0.73]
ISA3	[0.41, 2.11]	[−1.84, −0.12]	[−0.58, 1.20]
JPL1	[−0.36, 0.92]	[0.54, 2.03]	[−1.81, −0.43]
JPL2	[−0.34, 1.17]	[0.48, 2.37]	[−1.85, −0.07]
JPL3	[−0.52, 0.93]	[0.50, 2.28]	[−1.56, 0.25]
KHA1	[−1.18, 0.39]	[−3.07, −1.46]	[−1.19, 0.26]
KHA2	[−1.46, 0.15]	[−3.17, −1.32]	[−0.93, 0.61]
KHA3	[−1.71, 0.25]	[−2.95, −0.72]	[−1.25, 0.57]
LOT1	[−0.74, 0.61]	[−2.51, −0.87]	[−0.81, 0.61]
LOT2	[−0.69, 0.40]	[−1.94, −0.62]	[−0.80, 0.33]
LOT3	[−0.82, 0.34]	[−2.12, −0.70]	[−0.77, 0.52]
PHI1	[0.22, 1.51]	[0.56, 1.84]	[−1.40, −0.08]
PHI2	[−0.09, 1.66]	[0.33, 2.29]	[−1.81, 0.22]
PHI3	[−0.25, 1.38]	[0.25, 2.25]	[−2.01, −0.12]
ROM1	[2.19, 3.45]	[−1.20, 0.29]	[−0.51, 0.81]
ROM2	[1.85, 3.63]	[−1.30, 0.97]	[−0.83, 1.36]
ROM3	[1.48, 3.57]	[−1.33, 1.31]	[−0.79, 1.79]

obtain the second vertices principal component as PC2($\alpha = 0.2$) = [−2.051, −1.644].

Table 4 provides the complete set of vertices principal components obtained from Eq. (37) when $\alpha = 0.2$; these are plotted in Fig. 7. Also given in Table 4 are the numbers of vertices for which the contribution to the respective principal components ($\nu = 1, 2$) exceeds $\alpha = 0.2$. Under this criterion, seven of the observations now have a principal component for which all (here 64) vertices contribute less than $\alpha = 0.2$. In these cases, to anchor the (other) principal component, we take the average over the vertices. For example, for the face INC1 ($i = 7$) no

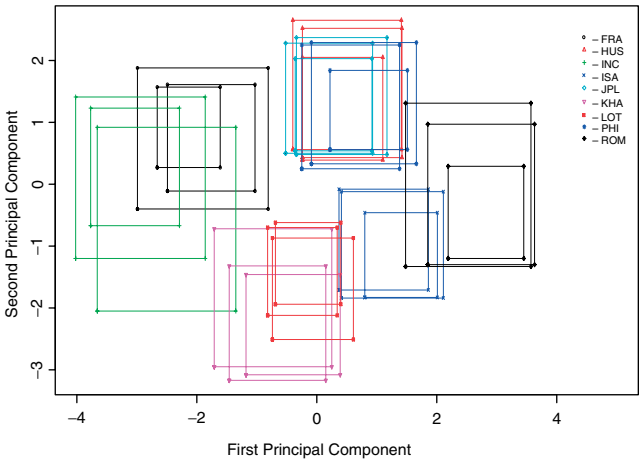

Fig. 6 Faces: vertices principal components PCν, $\nu = 1, 2$.

vertex contributes more than 0.2 to the second principal component. In this case, $\overline{y}_{7,2} = 0.28$ from Eq. (40). This is reflected as a line (instead of a rectangle) parallel to the first principal component axis in Fig. 7. Notice that the face JPL1 also assumes a linear form (parallel to the second principal component axis). In this case, however, this arises from Eq. (37) where now only one vertex contributes more than $\alpha = 0.2$ to the first principal component.

By comparing the principal components of Fig. 6 (i.e., $\alpha = 0.0$) and of Fig. 7 ($\alpha = 0.2$), the greater clarity of the classes is immediately apparent. Similar enhancements emerged as α moved from 0 to 0.6 (not shown). Specifically, four groups are evident, those containing the faces of $\{PHI, JPL, HUS\}$, $\{ROM, ISA\}$, $\{FRA, INC\}$ and $\{LOT, KHA\}$, respectively. An equivalent analysis using the second and three principal components PC2 and PC3 suggests this first group could be divided into two, $\{PHI, JPL\}$ and $\{HUS\}$.

Table 5 gives all the eigenvalues λ_ν, $\nu = 1, \ldots, 6$, along with the percentage and the cumulative percentage of the total variation explained by each principal component. Thus, we see that PC1 explains 42.7% of the total variation and the first two principal components (PC1 and PC2) together account for 72.7% of the total variation.

Table 3. Faces: vertices contributions to PC$\nu = 1, 2$ ($i = 8 \equiv$ INC2).

X_1	X_2	X_3	X_4	X_5	X_6	PC1	PC2	Cor1	Cor2
156.26	51.09	95.77	91.21	54.92	54.41	−2.717	−1.993	0.514	0.277
156.26	51.09	95.77	96.83	54.92	54.41	−2.390	−1.955	0.476	0.318
156.26	51.09	99.36	91.21	54.92	54.41	−2.511	−2.051	0.485	0.323
156.26	51.09	99.36	96.83	54.92	54.41	−2.184	−2.012	0.448	0.380
161.31	51.09	95.77	91.21	54.92	54.41	−2.432	−1.683	0.437	0.209
161.31	51.09	95.77	96.83	54.92	54.41	−2.105	−1.644	0.396	0.242
161.31	51.09	99.36	91.21	54.92	54.41	−2.226	−1.740	0.406	0.248
161.31	51.09	99.36	96.83	54.92	54.41	−1.899	−1.702	0.366	0.294

Table 4. Faces: vertices principal components, $\nu = 1, 2$, $\alpha = 0.2$.

	Principal component		# Vertices retained	
Subject	PC1	PC2	$\nu = 1$	$\nu = 2$
FRA1	[−2.66, −1.61]	[1.12, 1.57]	64	12
FRA2	[−2.49, −1.03]	[0.94, 1.61]	64	18
FRA3	[−2.99, −0.81]	[0.67, 1.87]	64	17
HUS1	[0.87, 1.10]	[0.81, 2.05]	3	49
HUS2	[0.86, 1.41]	[0.97, 2.65]	6	56
HUS3	[0.68, 1.42]	[0.88, 2.52]	11	50
INC1	[−3.77, −2.29]	0.28	64	0
INC2	[−3.66, −1.35]	[−2.05, −1.64]	64	8
INC3	[−4.02, −1.85]	0.11	64	0
ISA1	[0.80, 2.00]	[−1.83, −0.70]	64	51
ISA2	[0.67, 1.86]	[−1.71, −0.51]	52	38
ISA3	[0.66, 2.11]	[−1.84, −0.46]	60	41
JPL1	[0.92, 0.92]	[0.60, 2.03]	1	60
JPL2	[0.64, 1.17]	[0.79, 2.37]	7	57
JPL3	[0.81, 0.93]	[0.59, 2.28]	3	60
KHA1	−0.39	[−3.07, −1.46]	0	64
KHA2	[−1.46, −1.09]	[−3.17, −1.32]	4	64
KHA3	[−1.71, −0.83]	[−2.95, −0.72]	12	64
LOT1	−0.07	[−2.61, −0.87]	0	64
LOT2	−0.14	[−1.94, −0.62]	0	64
LOT3	−0.24	[−2.12, −0.70]	0	64
PHI1	[0.63, 1.51]	[0.63, 1.84]	36	59
PHI2	[0.62, 1.66]	[0.66, 2.29]	26	51
PHI3	[0.62, 1.38]	[0.62, 2.25]	18	54
ROM1	[2.19, 3.45]	−0.46	64	0
ROM2	[1.85, 3.63]	−0.17	64	0
ROM3	[1.48, 3.57]	[1.28, 1.31]	64	2

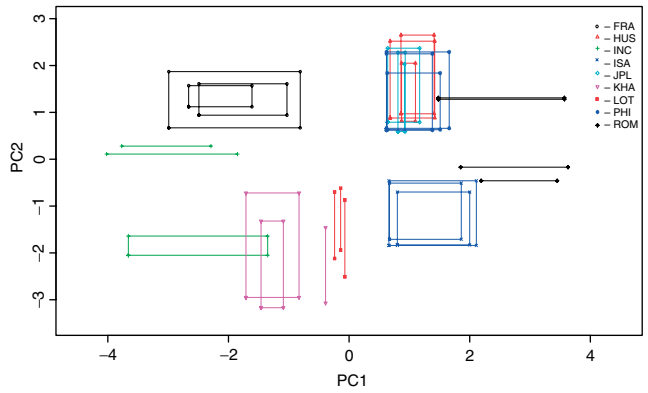

Fig. 7 Faces: vertices principal components PCν, $\nu = 1, 2$: $\alpha = 0.2$.

The correlations $C_{j\nu}$ between the variable X_j and the νth principal component PCν were calculated from Eq. (31) and are shown in Table 6, for $\nu = 1, 2, 3$. These suggest there is a strong relationship between the right and left distances and the upper middle lip ($X_3 = AH$ and $X_4 = DH$) and the first principal component PC1 with correlations of 0.84 and 0.89, respectively, followed by the eye-span distance ($X_1 = AD$) with a correlation of 0.64. These variables relate to the overall size of a face. The correlations of the variables with the second principal component PC2 reveal the relative importance of the interior facial detail, viz., the distance between the eyes $X_2 = BC$ has a correlation equal to 0.67; likewise $X_5 = EH$ and $X_6 = GH$ relating to the mouth with correlations of 0.62 and 0.76, respectively. Not surprisingly, it is the same set of variables which single out the faces $\{ROM, FRA, INC, ISA\}$ relative to the axis of PC1 from the other faces, and likewise those of $\{HUS, KHA, LOT\}$ relative to the axis of PC2 from the other faces; the details are omitted.

Table 5. Faces: vertices PC inertia.

PCν	Eigenvalue λ_ν	% Inertia	Cumulative inertia
PC1	2.560	42.7	42.7
PC2	1.798	30.0	72.7
PC3	0.642	10.7	83.4
PC4	0.476	7.9	91.3
PC5	0.335	5.6	96.9
PC6	0.188	3.1	100

Table 6. Faces: vertices method, correlations $C_{j\nu}$ between X_j and PCν.

X_j	PC1	PC2	PC3
AD	0.6444	0.5889	0.1717
BC	0.4903	0.6663	−0.1403
AH	0.8374	−0.1968	−0.3707
DH	0.8913	0.0885	0.1649
EH	−0.4749	0.6248	−0.5607
GH	−0.4283	0.7554	0.3377

On the basis of these results, we conclude that long faces (as in relatively long values of AH and DH) or oval shaped faces are projected into the positive plane of the first principal component, while the relatively rounder or broad faces are projected into the positive plane of the second principal component.

Further insights are obtained by studying the relative contributions $\mathrm{Ctr}(H_i, \mathrm{PC}_\nu)$ between the full observation $\boldsymbol{\xi}_i$ and the νth principal component. These values, calculated from Eq. (33), are given in Table 7 for $\nu = 1, 2, 3$. Thus, we observe that the faces of $\{FRA, INC, ISA, ROM\}$ are highly identified with the first principal component PC1, while those of $\{KHA, LOT\}$ have their highest contributions with the second principal component PC2. It becomes clear from the preceding discussion that $\{FRA, INC, ISA, ROM\}$ distinguish themselves through the importance of the (AH, DH and AD) variables, that is, they have long and/or oval faces. Characteristics of the other groups can likewise be identified.

Table 7. Faces: vertices method, relative contributions to vertices PCν, $\nu = 1, 2, 3$.

Subject	PC1	PC2	PC3
FRA1	0.70	0.14	0.04
FRA2	0.62	0.15	0.04
FRA3	0.64	0.14	0.05
HUS1	0.06	0.33	0.41
HUS2	0.07	0.45	0.32
HUS3	0.10	0.40	0.29
INC1	0.86	0.03	0.01
INC2	0.61	0.08	0.08
INC3	0.77	0.04	0.05
ISA1	0.51	0.33	0.02
ISA2	0.42	0.28	0.03
ISA3	0.44	0.27	0.07
JPL1	0.04	0.42	0.33
JPL2	0.08	0.43	0.21
JPL3	0.05	0.50	0.14
KHA1	0.04	0.78	0.05
KHA2	0.07	0.77	0.03
KHA3	0.12	0.60	0.06
LOT1	0.02	0.68	0.04
LOT2	0.02	0.54	0.05
LOT3	0.03	0.53	0.05
PHI1	0.24	0.41	0.16
PHI2	0.21	0.43	0.19
PHI3	0.14	0.37	0.29
ROM1	0.86	0.04	0.01
ROM2	0.83	0.04	0.05
ROM3	0.73	0.06	0.08

These conclusions are based on using all the vertices for a given hypercube as a collective whole. If we return to the contributions of individual vertices and in particular those that exceed α (=0.2, in Table 4), our conclusions are further corroborated and strengthened. For example, take the faces of LOT (as an extreme case). From Table 7, we observe that the contributions to the second principal component are the largest over all faces at 0.68, 0.54, 0.53, respectively, while those to the first principal component are the smallest of all faces at 0.02, 0.02, 0.03, respectively. When we considered individual vertices, all 64 vertices were retained for the second principal component whereas none were retained for the first principal component. At the other extreme, we have the faces of ROM with strong contributions to the first principal component both collectively as a complete hypercube and individually as vertices; in this case, the contributions to the second principal component are weak. Likewise, enhanced interpretations apply to the other faces, with the faces of ISA being a 'central' face balanced over both principal components. Notice, from Fig. 7 that the ISA faces essentially form their own cluster.

Finally, in Table 8, in the first three columns, we provide the contribution $I_{i\nu}$ of the variance λ_ν of the principal component PCν, $\nu = 1, 2, 3$, for each observation, obtained from Eq. (34). Then, in the right-hand column, we give this contribution I_i of each observation to the total variance, calculated from Eq. (35). Thus, for example, we observe that the faces INC($I_{i1} = 0.13, 0.09, 0.13$) and ROM($I_{i1} = 0.12, 0.11, 0.09$) contribute the most variation to λ_1. The same faces INC($I_i = 0.07, 0.06, 0.07$) closely followed by ROM($I_i = 0.06, 0.06, 0.06$) contribute most to the overall variation.

5.3. Classical Surrogates

In the absence of any methodology for interval-valued data, it would be necessary to adopt a classical surrogate for the symbolic data; three are considered. The results are then compared with the symbolic analysis, from which it becomes evident that the classical analyses are unable to capture all the information contained in the original symbolic data.

One surrogate is the midpoint value obtained by replacing the symbolic interval $x = [a, b]$ by its classical midpoint $z = (a + b)/2$. A standard principal component analysis can then be conducted on the resulting $m \times p$ $(= 27 \times 6)$

Table 8. Faces: vertices method, absolute contributions of subject to PCν and inertia.

Subject	PC1	PC2	PC3	Inertia
FRA1	0.07	0.02	0.01	0.04
FRA2	0.05	0.02	0.01	0.03
FRA3	0.06	0.02	0.01	0.04
HUS1	0.00	0.03	0.12	0.03
HUS2	0.01	0.06	0.11	0.04
HUS3	0.01	0.05	0.10	0.04
INC1	0.13	0.01	0.01	0.07
INC2	0.09	0.02	0.05	0.06
INC3	0.13	0.01	0.03	0.07
ISA1	0.03	0.03	0.00	0.03
ISA2	0.02	0.02	0.01	0.02
ISA3	0.02	0.02	0.02	0.02
JPL1	0.00	0.04	0.08	0.03
JPL2	0.00	0.05	0.07	0.03
JPL3	0.00	0.04	0.04	0.03
KHA1	0.00	0.11	0.02	0.04
KHA2	0.01	0.11	0.01	0.04
KHA3	0.01	0.08	0.02	0.04
LOT1	0.00	0.06	0.01	0.03
LOT2	0.00	0.04	0.01	0.02
LOT3	0.00	0.04	0.01	0.02
PHI1	0.01	0.03	0.04	0.02
PHI2	0.01	0.04	0.05	0.03
PHI3	0.01	0.04	0.08	0.03
ROM1	0.12	0.01	0.01	0.06
ROM2	0.11	0.01	0.03	0.06
ROM3	0.09	0.01	0.04	0.06

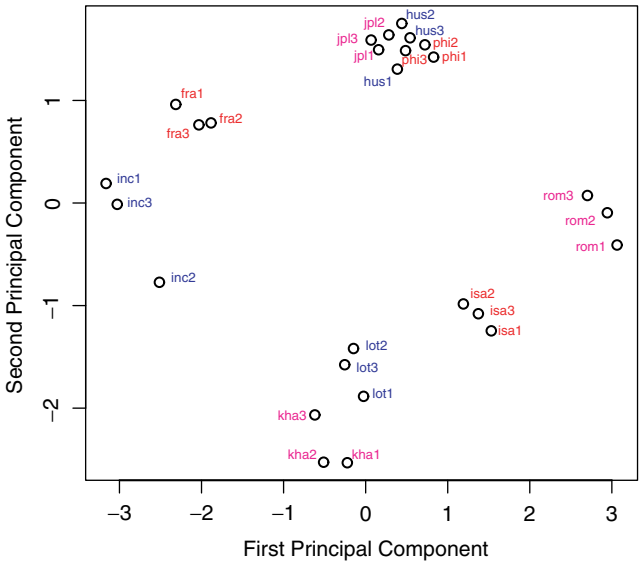

Fig. 8 Faces: classical principal components PCν, $\nu = 1, 2$ - Midpoints.

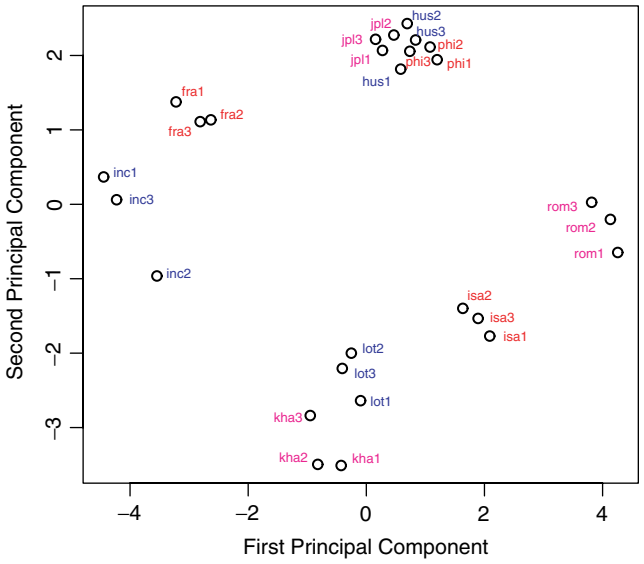

Fig. 9 Faces: classical principal components PCν, $\nu = 1, 2$ - Endpoints.

classical dataset. A plot of the first and second principal components for these data is shown in Fig. 8. One limiting factor of this surrogate is that it is impossible to retain any measure of the internal variation; for example, the two intervals $x_1 = [155, 157]$ and $x_1^* = [150, 163]$ both give the same surrogate $z_1 = 156$. It is not possible for this classical analysis to capture the difference between these two intervals.

Therefore, a possible way to overcome this limitation is to introduce two surrogate variables for each interval variable, viz., the interval endpoints. That is, the symbolic interval variable $x = [a, b]$ is replaced by $z_1 = a$ and $z_2 = b$. Then, a standard principal component analysis can be performed on the resulting $m \times 2p$ $(= 27 \times 12$ here) classical dataset. Figure 9 shows the plot of the first and second principal component analysis that ensues.

Except for the scale of the principal components, these two surrogates produce remarkably similar results. As for the symbolic analysis, the coherency of the three observations relating to each of the nine faces is evident. Four groups emerge, viz., those containing the faces of {PHI, JPL, HUS}, {FRA, INC}, {ISA, HA, LOT}, and {ROM}, though it can be argued that the second group should be broken into the individual faces {FRA} and {INC}, and the third group into {KHA, LOT} and {ISA}.

Rather than the endpoints, another possible way to accommodate intervals of differing lengths is to replace the interval variable by two variables, viz., the midpoint and range variables, such as used by Giordani and Kiers [17] and Lauro and Palumbo [5]. Now, the symbolic interval $x = [a, b]$ is replaced by $z_1 = (a + b)/2$ and $z_2 = (b - a)$. Then, as for the previous two surrogates, a standard classical analysis is conducted on the resulting $m \times 2p$ $(= 27 \times 12)$ classical dataset. The plot of the first and second principal components is shown in Fig. 10.

These three surrogate analyses are compared through Figs. 8–10. While both Figs. 8 and 9 retain the coherences for the sets of the same three faces observed for the symbolic analysis, the range surrogate in general loses that coherence (though it is seen in some cases, e.g., faces LOT, and KHA albeit to a lesser extent, i.e., the coherence is not as strong). This is particularly evident when comparing Fig. 8 for the midpoints with Fig. 10 when ranges are also

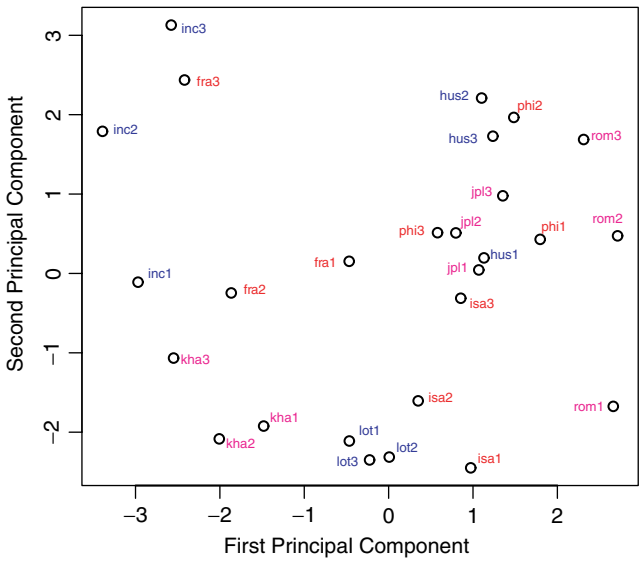

Fig. 10 Faces: classical principal components PCν, $\nu = 1, 2$—ranges and midpoints.

used along with the midpoint variable. It is also clear from Fig. 10 that clusters are mixtures of faces, for example, faces {INC1, INC3, FRA3} could be one cluster. One explanation for these incoherencies may be the fact that two intervals such as $x = [1, 11]$ and $x^* = [101, 111]$, say, with the same range (=10) but very different locations (6 and 106, respectively) are indistinguishable when the range variable has a high correlation with a PC value. This is akin to the problem of two different intervals but with the same midpoints giving the same results for an analysis based on midpoints.

5.4. Comparison of Symbolic and Surrogate Analyses

For comparison purposes, consider the vertices principal component analysis and the midpoint surrogate analysis; similar conclusions pertain when the endpoints surrogate analyses are included. However, given the fact that the ranges surrogates are inconclusive and inconsistent, no further comparison of those surrogates will be made.

The major difference between these analyses is that the vertices results reflect all the variations between the observations including internal variations, whereas the classical results do not. The classical values plotted in Fig. 8 are points in space, while the symbolic values are hypercubes, here rectangles for the $s = 2$ principal components plotted in Fig. 6. These rectangles have smaller (or larger) dimensions whenever the original data are smaller (or larger) intervals. Compare, for example, ROM1 and ROM3. Here, ROM1 has a smaller PC1 $\times$ PC2 rectangle, reflecting the smaller X_5 and X_6 intervals. Superimposing rectangles express a similarity between the corresponding face prototypes, and the size of the rectangle conveys the amount of variability through the corresponding 27 acquired face images. That is, the principal components themselves reflect a measure of internal variations along with a measure of the variation between observations. The classical analysis can only detect measures of the between observation variations, and as such do not reflect all the variations in the data. Notice that even the endpoint surrogate analysis fails to identify these internal variations in the final principal components (compare Figs. 6 and 9).

For these data, the two analyses produce slightly different groups of faces (though arguments can be made for the same groupings). In particular, the classical analysis suggests the ISA face belongs in the same group as the KHA and LOT faces, whereas the symbolic analysis suggests the ISA face is grouped with the ROM face. Certainly, in Fig. 6, this ISA face has its principal component region overlapping those of ROM and largely disjoint from the LOT and KHA regions. This distinction becomes more pronounced when $\alpha = 0.2$ (see Fig. 7), where it is more obvious that the ISA and ROM faces belong to the same group. Notice too that, from Fig. 7 the ROM3 face is in closer proximity to the {HUS, JPL, PHI} group (at least relative to the second principal component) than to its namesake group {ROM1, ROM2, ISA}. It also follows from Fig. 7 and Tables 6 and 7 that the faces LOT express their internal variation almost entirely through the variables AD, BC, GH, and EH (i.e., on the eyes and the mouth) and not at all on AH and DH (i.e., the distances from the eyes to the mouth); while in contrast the faces INC (and also ROM) are such that their internal variations are characterized by the eyes to mouth distances AH and DH and not at all by the eyes and mouth variables (AD, BC, GH, and EH). Such insight and information can not be educed from the classical analysis. The types of clarifications that can emerge for $\alpha > 0$ in a symbolic analysis are not possible in a classical approach. These differences in conclusions are a direct result of the fact that symbolic analyses are able to incorporate internal variations in the data into the methodology, thus enhancing the interpretations and expanding the knowledge gained.

6. WINE TASTERS DATASET

For standard datasets, dimensionality problems exist when the number of observations m is less than p the number of variables. However, it is not a problem for the vertices symbolic methodology unless $n < p$, where n is the total number of vertices given in Eq. (2). If there are no trivial intervals in the dataset, this becomes $n = m2^p < p$.

This is illustrated by performing the vertices principal component analysis on the data of Table 9. These data are the first $m = 4$ tasters of the first $p = 6$ different wines, extracted from a larger dataset of 23 wines and 21 wine tasters. Here, $n = 4 \times 2^6 = 256 > 6 = p$; so the methodology works through routinely. Each of the observations relates to a wine taster who makes an assessment of taste on each wine.

Further, the interval taste valuation indicates the extent of the taster's uncertainty of the quality of that wine, with a smaller range representing greater certainty as to its quality. This suggests that these varying uncertainties are best accommodated by using inverse weights such as Eq. (8). In this way, tasters who exhibit a greater uncertainty on their evaluation are penalized, while values for the more certain tasters contribute more to the determination of the principal component axes.

Therefore, by inserting the weights of Eq. (8) into the methodology, the first and second weighted vertices principal components are as shown in Table 10, and plotted in Fig. 11. Again, as for the previous analysis, the sizes of the principal component hypercubes (here rectangles)

Table 9. Wine tasters.

	Wine 1	Wine 2	Wine 3	Wine 4	Wine 5	Wine 6
Taster 1	[56, 74]	[75, 92]	[83, 90]	[47, 82]	[68, 86]	[42, 90]
Taster 2	[83, 85]	[89, 94]	[83, 89]	[48, 87]	[84, 91]	[81, 91]
Taster 3	[84, 90]	[86, 92]	[87, 93]	[82, 86]	[88, 95]	[86, 90]
Taster 4	[80, 91]	[85, 93]	[85, 92]	[73, 76]	[85, 92]	[81, 91]

reflect the relative sizes of the original data. Thus, taster 1 has a wider range of uncertainty than did taster 2 (except for wine 4) and this is reflected by a larger principal component rectangle. It is also evident that taster 1 has different taste evaluations overall than do the other three tasters who are more consistent with each other.

Calculating the principal components for $\alpha > 0$ can also be done; but for these data, the two groups $G_1 = \{\text{taster } 1\}$ and $G_2 = \{\text{tasters } 2, 3, 4\}$ are already visually distinct when $\alpha = 0$. However, the principal component rectangles when $\alpha = 0.2$ are also shown in Fig. 11, indicated by the dashed lines. These rectangles are unchanged for tasters 2 and 4 and that for taster 3 is reduced in size; however, it is still clear that these three tasters belong to a single group. Taster 1 is now more obviously a separate identity. The vertices in the analysis for which this contribution to the second principal component is less than $\alpha = 0.2$ revolve around the sixth wine Y_6. All vertices for which $Y_6 = 42$ are retained in Eq. (37) along with a few (four) vertices with $Y_6 = 90$. That is, vertices with $Y_6 = 42$ are important in explaining the underlying variations of the data, whereas vertices with $Y_6 = 90$ are less important. This in effect corroborates the earlier conclusions from the relative sizes of the principal component rectangles that taster 1 is considerably more uncertain than are the other three tasters, and further that it is the lower levels of the uncertainty interval that distinguishes taster 1, most especially on wine 6. These conclusions are even more pronounced when $\alpha = 0.6$ (see Fig. 11).

The correlations between the wines and principal components ($\nu = 1, 2, 3$) are shown in Table 11. Interpretation is left to the reader, other than to comment that wines 1, 2, and 6 dominate the distinctions between the tasters. Table 12 provides the cumulative variations by principal components. Thus, the first three principal components account for 60% of the variation.

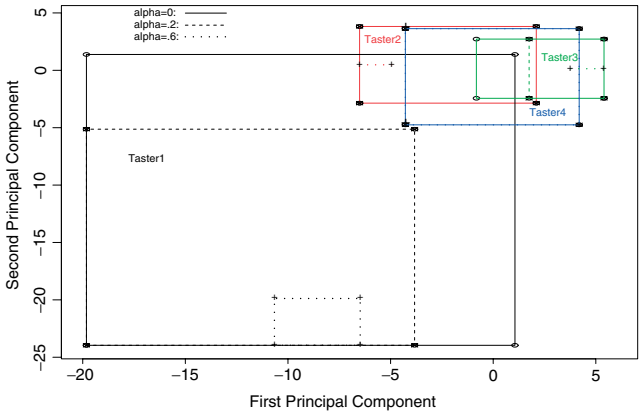

Fig. 11 Wines: weighted vertices principal components, $\alpha = 0.0, 0.2, 0.6$.

Table 11. Wines: correlations.

X_j	PC1	PC2	PC3
Wine 1	0.517	0.197	0.056
Wine 2	0.305	0.623	0.176
Wine 3	0.274	−0.196	−0.055
Wine 4	0.264	−0.072	−0.020
Wine 5	0.247	0.124	0.035
Wine 6	0.233	0.712	0.201

Table 12. Wines: PC variation.

	λ_j	Variation	Cumulative variation
PC1	1.687	28.1	28.1
PC2	0.994	16.6	44.7
PC3	0.894	14.9	59.6
PC4	0.860	14.3	73.9
PC5	0.805	13.4	87.4
PC6	0.759	12.7	100

7. CONCLUSION

Symbolic data emerge in numerous ways in contemporary datasets. This work has focused on principal component methodology for interval-valued data. In particular, enhancements were made to the vertices method allowing

Table 10. Wines: weighted vertices principal components.

	$\alpha = 0.0$		$\alpha = 0.2$		Number of vertices	
	PC1	PC2	PC1	PC2	$\nu = 1$	$\nu = 2$
Taster 1	[−19.816, 1.065]	[−23.960, 1.380]	[−19.816, −3.834]	[−23.960, −5.133]	38	36
Taster 2	[−6.506, 2.096]	[−2.862, 3.817]	[−6.506, 2.096]	[−2.862, 3.817]	32	19
Taster 3	[−0.824, 5.404]	[−2.444, 2.723]	[1.745, 5.404]	[−2.444, 2.723]	43	12
Taster 4	[−4.275, 4.186]	[−4.750, 3.626]	[−4.275, 4.186]	[−4.750, 3.626]	17	26

for trivial intervals, weights, constrained observations, and visualizations involving the vertices of the data hypercubes, concepts not considered by other interval methods (most of which usually involved interval midpoints in some fashion). Compared to other available methods, the vertices method proved superior most especially with respect to computational complexity and optimum covering envelopes issues. Further, analyses using classical surrogates produced results that failed to capture all the variation inherent to the data. For example, a classical analysis using the interval midpoints ignores internal variations present in the data; the centers method has the same limitation. Methods involving a range variable could not always distinguish between differing observations with similar range values.

An alternative analysis of the faces data could be to establish the five-number summaries (minimum, first quartile, median, third quartile, maximum) introduced by Tukey [40]. This produces a set of categorical valued observations. Then, the Ichino [41] approach could be applied to the resulting dataset.

In addition to intervals, different aggregations of large databases guided by different scientific questions could instead produce symbolic datasets consisting of other forms of symbolic data such as lists, or modal-valued observations, for example, histograms or probability distributions. Given the inevitable continued growth in the size of datasets, it is inevitable that such datasets will become 'routine'. Therefore, it is important to develop principal component methodologies for other classes of symbolic data such as multivalued and histogram-valued data. There is also a need to develop theoretical underpinnings to all these methods. These remain as outstanding problems for future researchers.

Finally, algorithms for executing vertices principal component analyses for interval data are available at http://www.ceremade.dauphine.fr/touati/sodaspagegarde.htm, the SODAS1.4 webpage.

ACKNOWLEDGMENT

Partial funding support from the National Science Foundation is gratefully acknowledged.

REFERENCES

[1] I. T. Jolliffe, Principal Component Analysis, New York, Springer-Verlag, 1986.
[2] L. Billard, and E. Diday, From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis, J Am Stat Assoc 98 (2003), 470–487.
[3] L. Billard, and E. Diday, Symbolic Data Analysis: Conceptual Statistics and Data Mining, Chichester, John Wiley, 2006.
[4] P. Cazes, A. Chouakria, E. Diday, and Y. Schechtman, Extension de l'analyse en composantes principales à des données de type intervalle, Rev Stat Appl 45 (1997), 5–24.
[5] C. N. Lauro, and F. Palumbo, Principal component analysis of interval data: a symbolic analysis approach, Comput Stat 15 (2000), 73–87.
[6] A. Chouakria, Extension des methodes d'analyse factorielle a des donnees de type intervalle, These de doctorat. University Paris Dauphine, 1998.
[7] F. Palumbo, and C. N. Lauro, A PCA for interval valued data based on midpoints and radii, In New Developments in Psychometrics, H. Yanai, A. Okada, K. Shigematu, Y. Kano, and J. J. Meulman, eds. Japan, Springer-Verlag, 2003, 641–648.
[8] C. N. Lauro, and F. Palumbo, Principal component analysis for non-precise data, In New Developments in Classification and Data, M. Vichi, P. Morani, S. Mignami, and A. Montanari, eds. Berlin, Springer-Verlag, 2005, 173–183.
[9] R. E. Moore, Interval Analysis, New Jersey, Prentice Hall, 1966.
[10] F. Gioia, and C. N. Lauro, Principal component analysis on interval data, Comput Stat 21 (2006), 343–363.
[11] C. N. Lauro, and F. Gioia, Dependence and interdependence analysis for interval-valued variables, In Data Science and Classification, V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, eds. Berlin, Springer-Verlag, 2006, 171–183.
[12] A. S. Deif, Singular values of an interval matrix, Linear Algebra Appl 151 (1991), 125–133.
[13] J. Rhon, Interval matrices: singularity and real eigenvalues, SIAM J Matrix Anal Appl 14 (1993), 82–91.
[14] P. D'Urso, and P. Giordani, A least squares approach to principal component analysis for interval valued data, Chem Intell Lab Syst 70 (2004), 179–192.
[15] T. Denoeux, and M. H. Masson, Principal component analysis of fuzzy data using autoassociative neural networks, IEEE Trans Fuzzy Syst 12 (2004), 336–349.
[16] P. Giordani, and H. A. L. Kiers, Principal component analysis of symmetric fuzzy data, Comput Stat Data Anal 45 (2004), 519–548.
[17] P. Giordani, and H. A. L. Kiers, A comparison of three methods for principal component analysis for fuzzy interval data, Comput Stat Data Anal 51 (2006), 379–397.
[18] R. Coppi, P. Giordani, and P. D'Urso, Component models for fuzzy data, Psychometrika 71 (2006), 733–761.
[19] Y. Yabuuchi, J. Watada, and Y. Nakamori, Fuzzy principal component analysis for fuzzy data, In: Proceedings Sixth IEEE International Conference on Fuzzy Systems 2, 1997, 1127–1132.
[20] B. Leroy, A. Chouakria, I. Herlin, and E. Diday, Approche géométrique et classification pour la reconnaissance de visage, Reconnaissance des Forms et Intelligence Artificelle, INRIA and IRISA and CNRS, France, 1996, 548–557.
[21] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, Face recognition: a literature survey, ACM Comput Surv 35 (2003), 399–459.
[22] F. A. T. De Carvalho, Extension based proximity coefficients between Boolean symbolic objects, In Data Science, Classification, and Related Methods, C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, and Y. Baba, eds. Berlin, Springer-Verlag, 1998, 370–378.
[23] R. A. Johnson, and D. W. Wichern, Applied Multivariate Statistical Analysis (5th ed.), New Jersey, Prentice Hall, 2002.

[24] T. W. Anderson, An Introduction to Multivariate Statistical Analysis (2nd ed.), New York, John Wiley, 1984.
[25] M. A. Fischler, and R. A. Eschlager, The representation and matching of pictorial structures, IEEE Trans Comput c-22 (1973), 67–92.
[26] R. J. Baron, Mechanisms of human facial recognition, Int J Man Mach Stud 15 (1981), 137–178.
[27] A. Samal, and P. Iyengar, Automatic recognition and analysis of human faces and facial expressions. A survey, Pattern Recogn 25 (1992), 65–77.
[28] R. Chellappa, C. L. Wilson, and S. Sirohey, Human and machine recognition of faces, a survey, Proc IEEE 83 (1995), 705–740.
[29] M. Turk, and A. Pentland, Eigenfaces for recognition, J Cogn Neurosci 3 (1991), 72–86.
[30] I. Craw, and P. Cameron, Face recognition by computer, In Proceedings British Machine Vision Conference, 1996. 489–507.
[31] H. Moon, and P. J. Phillips, Computational and performance aspects of PCA-based face recognition algorithms, Perception 30 (2001), 301–321.
[32] K. Etemad, and R. Chellappa, Discriminant analysis for recognition of human fac images, J Opt Soc Am 14 (1997), 1724–1733.
[33] B. Moghaddam, and A. Pentland, Probabilistic visual learning for object representation, IEEE Trans Pattern Anal Mach Intell 19 (1997), 696–710.
[34] B. Li, and R. Chellappa, A generic approach to simultaneous tracking and verification in video, IEEE Trans Image Process 11 (2002), 530–544.
[35] S. Z. Li, and J. Lu, Face recognition using the nearest feature line method, IEEE Trans Neural Networks 10 (1999), 439–443.
[36] M. Kass, A. Witkin, and D. Terzopoulos, Snakes: active contour models, In IEEE Proceedings of the International Conference on Computer Vision, 1987, 259–268.
[37] M. Turk, Interactive-time vision: face recognition as a visual behavior, PhD Thesis, Massachusetts Institute of Technology, 1991.
[38] I. Craw, D. Tock, and A. Bennett, Finding face features, In Proceedings of the Second European Conference on Computer Vision, 1992, 92–96.
[39] L. H. Staib, and J. S. Duncan, Boundary finding with parametrically deformable models, IEEE Trans Pattern Anal Mach Intell 14 (1992), 1061–1075.
[40] J. W. Tukey, Exploratory Data Analysis, Reading, MA, Addison-Wesley, 1977.
[41] M. Ichino, The quantile method for symbolic principal component analysis, Stat Anal Data Mining (2011), 4, in press.