Symbolic principal component for interval-valued observations

Lynn Billard, Ahlame Douzal-Chouakria, Edwin Diday

SYMBOLIC PRINCIPAL COMPONENTS FOR INTERVAL-VALUED OBSERVATIONS

L. Billard[1], A. Douzal-Chouakria[2], E. Diday[3]
[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA
[2]TIMC-IMAG, Université Joseph Fourier Grenoble 1,
IN3S INstitute de l'INgénierie de l'INformation de Santé,
Facult de Médecine, F-38706 LA TRONCHE Cedex, France
[3]Ceremade, University of Paris Dauphine, 75775 Paris Cedex 16 France

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 extends the methodology of classical principal components to that for interval-valued data. Two methods are proposed, viz., a vertices method which uses all the vertices of the observation's hypercube, and a centers method which uses the centroid values. Unlike classical data, each symbolic data point has internal variation. For both the vertices and centers methods, we obtain interval-valued symbolic principal components which recapture the internal variation of the observations, as well as diagnostics such as correlation measures between these principal components and the random variables and/or the observations themselves. We also provide a visualization method that further aids in the interpretation of the methodology. The methods are illustrated in a dataset using measurements of facial characteristics obtained from a study of face recognition patterns for surveillance purposes, and in a dataset of species of bats where the measurements are naturally internal-valued. A comparison with analyses in which classical surrogates replace the intervals, shows how the symbolic analyses give more informative conclusions.

Keywords: Vertices principal components, centers principal components, 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 $s << p$, 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 $\Re^p$; see, e.g., Jolliffe (1986). The goal of this paper is to describe principal component methodology for interval-valued symbolic data.

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. For example, consider a health insurance dataset of millions of observations relating to individuals covered by an insurer. Each entry may record name, dates visited, age, weight, ... (other demographic and geographic information) ..., pulse rate, blood pressure, ... (other basic medical diagnostics) ..., lists of ailments, lists of diagnoses, lists of treatments prescribed, ..., and so on. Rather than the details of a particular individual on a specific hospital visit, the insurer may be more interested in the medical patterns of 20-year old females (say), or more generally in age × gender classifications. Aggregating the entire dataset over age × gender will produce values for weight (say) such as: 107, 98, 149, ...; that is, now the variable weight for the "observation" corresponding to 20-year old females takes values in the interval [98, 149], say. There are as many ways of aggregating the data as there are questions of interest. Thus, instead of age × gender, the question could be "Lung cancer patients × City of residence", age alone, heart patients × product treatment, and so on. The resulting database would now contain symbolic interval-valued data. [Some aggregations produce other types of symbolic data, such as lists or modal-valued observations, e.g., histograms or probability distributions; these are not considered herein.]

Interval-valued data can arise in their own right, as in, e.g., the oils dataset of Ichino (1988) used as an example in Chouakria et al. (2000). Another example relates to (say) characteristics of a species; e.g., the bat species *Pipistrelle Commune (Pipistrellus pipistrellus)* has height from 4 to 7 mm (i.e., the interval [4,7]) but a particular bat may have a height of 4.3mm. 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 (wine quality, e.g.) 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 say, e.g., that our pulse rate is 64 ± 1, i.e., [63, 65], or that our weight fluctuates between 60 and 62 kilos, i.e., [60, 62]. Note however that weights of 60 ± 1 and 60 ± 3 whilst having the same midpoint have different internal variations, 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, i.e., variance. A review of symbolic data can be found in Bock and Diday (2000) and Billard and Diday (2003, 2006).

The problem of reducing a large number of random variables p to a smaller number of principal components $s << p$ remains regardless of how the intervals were formed. Therefore, in Section 3 and 4, two methods for performing a principal component analysis on interval-valued data are presented, the vertices method and the centers method, respectively; practical considerations and computational complexity along with a discussion of the relative merits of these methods are presented in Section 5. These methods are based on extending the methodology for classical data, and we show how the basic classical theory carries through to interval-valued data. We also in Section 6 extend classical concepts relating to the interpretation of the principal components, such as inertia and contributions from an observation to a principal component. Unlike classical data which consist of single points in p-dimensional space, symbolic interval data consist of hypercubes and in particular each observation, i.e., each hypercube, consists 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 symbolic principal components can therefore be further enhanced by focusing on those vertices whose contributions exceed pre-assigned bounds; see Section 6. Section 7 considers constrained hypercubes. These constraints could arise naturally as part of an aggregation process behind the formation of some symbolic datasets. We start with some basic principles including structuring the symbolic data so as to retain the fundamental internal variations inherent to such data, along with some generalized weighting schemes for interval data, in Section 2.

In Section 8, we apply these methods, with $p = 6$ variables, to a set of $m = 27$ faces dataset from Leroy et al. (1996) 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. (2003) 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 new 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 analyse interval-valued data with classically valued surrogates lose information contained in the data; i.e., the symbolic analysis gives more informative results than does a classical analysis, lending more importance to the usefulness of the symbolic approach.

The faces dataset considered in Section 8 arose after aggregation of a much larger dataset. In contrast, in Section 9, we analyse a dataset with $p = 4$ physical characteristics (e.g., height) of $m = 21$ species of bats, data that are naturally interval-valued. The symbolic principal component analysis proposed herein reveals features of this dataset not possible to highlight from (surrogate) classical analyses. These datasets are new, providing additional insights from symbolic analyses beyond those of previous works that focussed on the much smaller ($m = 8$) oils dataset of Ichino(1988).

A preliminary version of the proposed methods was reported in conference proceedings by Chouakria et al. (1995, 1998). More complete details are in the doctoral dissertation Chouakria (1998, Ch. 1), a summary of which is in Cazes et al. (1997). Later, in an attempt to improve the factorial visualization of symbolic observations, Lauro and Palumbo (2000) considered three variants, each based on the interval midpoints (a special case of our proposed centers method). In a different direction, Palumbo and Lauro (2003) and Lauro and Palumbo (2005) use interval arithmetic ideas (of Moore, 1966) to calculate the variance-covariance matrix based on interval means and distances. Gioia and Lauro (2006) and Lauro and Gioia (2006) 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 (1991) and Rhon(1993). There is also a series of papers (see, e.g., D'Urso and Giordani, 2004, Denoeux and Masson, 2004, Giordani and Kiers, 2004, 2006, Coppi et al., 2006, and Yabuuchi et al., 2007) 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 Billard and Diday (2006) for examples showing the distinctions between these two types of data.

2 Preliminary Results

2.1 Data Interval Coding

Suppose the data consist of m observations $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})$ where

$$\xi_{ij} = [a_{ij}, b_{ij}], \quad i = 1, \ldots, m, \quad j = 1, \ldots, p,$$

with $a_{ij} \leq b_{ij}$, as 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

$$n_i = 2^{q_i}. \tag{2.1}$$

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. 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

$$n = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} 2^{q_i}. \tag{2.2}$$

We construct the data matrix $\boldsymbol{X}_{\xi_i}$ with elements

$$\boldsymbol{X}_{\xi_i} = \begin{pmatrix} x_{11}^i & \cdots & x_{1p}^i \\ \vdots & & \vdots \\ x_{k1}^i & \cdots & x_{kp}^i \\ \vdots & & \vdots \\ x_{n_i1}^i & \cdots & x_{n_ip}^i \end{pmatrix} \tag{2.3}$$

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 $\boldsymbol{\xi}_i$, $i = 1, \ldots, m$. For example, if an observation

$$\boldsymbol{\xi}_i = ([a_{i1}, b_{i1}], \quad [a_{i2}, b_{i2}], \quad [a_{i3}, b_{i3}])$$

with $a_{i3} = b_{i3}$, then $q_i = 2, n_i = 4$, and hence

$$\boldsymbol{X}_{\xi_i} = \begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \\ a_{i1} & b_{i2} & a_{i3} \\ b_{i1} & a_{i2} & a_{i3} \\ b_{i1} & b_{i2} & a_{i3} \end{pmatrix}.$$

The data matrix whose elements represent the vertices of the complete dataset is the $n \times p$ matrix, from (2.3),

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{\xi_1} \\ \vdots \\ \boldsymbol{X}_{\xi_m} \end{pmatrix} = \left(\begin{array}{c} \begin{pmatrix} x_{11}^1 & \cdots & x_{1p}^1 \\ \vdots & & \vdots \\ x_{n_1 1}^1 & \cdots & x_{n_1 p}^1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{11}^m & \cdots & x_{1p}^m \\ \vdots & & \vdots \\ x_{n_m 1}^m & \cdots & x_{n_m p}^m \end{pmatrix} \end{array} \right). \tag{2.4}$$

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.

An alternative coding is one which replaces the interval values $\boldsymbol{\xi}_i$ for each observation by the center-values $\boldsymbol{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)$ where, e.g.,

$$x_{ij}^c = (a_{ij} + b_{ij})/2. \tag{2.5}$$

Then, the associated data matrix is

$$\boldsymbol{X}^c = \begin{pmatrix} x_{11}^c & \ldots & x_{1p}^c \\ \vdots & & \vdots \\ x_{m1}^c & \ldots & x_{mp}^c \end{pmatrix}. \tag{2.6}$$

Therefore, instead of the m clusters of n total vertices as in Figure 1, we now have m centroids $\boldsymbol{x}_i^c$, $i = 1, \ldots, m$, points in $\Re^p$.

In the sequel, we develop two methods, one based on the data matrix $\boldsymbol{X}$ of (2.4) called the vertices method, and one based on the data matrix $\boldsymbol{X}^c$ of (2.6) called the centers method.

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

$$w_i = \sum_{k=1}^{n_i} w_k^i, \qquad \sum_{i=1}^{m} w_i = 1. \tag{2.7}$$

A frequent choice of weight for w_i gives equal weight to all observations, i.e.,

$$w_i = 1/m, \quad i = 1, \ldots, m. \tag{2.8}$$

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 (2.8).

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

$$w_i = V_i / \sum_{i=1}^{m} V_i, \tag{2.9}$$

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

$$V_i = \prod_{a_{ij} \neq b_{ij}} (b_{ij} - a_{ij}). \tag{2.10}$$

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 $\boldsymbol{\xi}_i$ (DeCarvalho,1998, and Lauro and Palumbo, 2000).

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

$$w_i = \frac{1 - V_i / \sum_{i=1}^{m} V_i}{\sum_{i=1}^{m} [1 - V_i / \sum_{i=1}^{m} V_i]}. \tag{2.11}$$

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. When all observations are classical, the inverse weights of (2.12) reduce to $w_i = 1/m$, $i = 1, \ldots, m$, of (2.8) so recapturing the traditional analysis as a special case.

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,

$$w_k^i = w_i/n_i, \quad k = 1, \ldots, n_i, \quad i = 1, \ldots, m. \tag{2.12}$$

Clearly, (2.7) is satisfied, since

$$\sum_{k=1}^{n_i} w_k^i = \sum_{k=1}^{n_i} \frac{w_i}{n_i} = w_i.$$

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 (2.1) $n_1 = 8$ 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 x_{ij}^c located at the midpoint values, as given in (2.5), which in effect is assuming a uniform distribution within the intervals.

More generally for any distribution, rather then the centroid value x_{ij}^c, 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 end-points a_{ij} and b_{ij}, respectively, such that

$$w_{ij}^a + w_{ij}^b = 1 \tag{2.13}$$

and

$$w_{ij}^a a_{ij} + w_{ij}^b b_{ij} = x_{ij}^0. \tag{2.14}$$

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

$$w_k^i = w_i \left[\prod_{j=1}^{q_i} w(x_{kj}^i) \right] \tag{2.15}$$

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

$$w(x_{kj}^i) = w_{ij}^t, \text{ when } x_{kj} = t_{ij},\ t = a, b. \tag{2.16}$$

It is easily verified, from (2.14), that (2.7) is satisfied.

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

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

Then, after applying (2.13) 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 midpoint of (2.5), the weights $w_{ij}^a = w_{ij}^b = 1/2$, and so the particular weights of (2.12) pertain.

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

$$\boldsymbol{D} = diag(w_1^1, \dots, w_{n_1}^1, \dots, w_1^m, \dots, w_{n_m}^m). \tag{2.17}$$

Classical Data:

When all the observations are classical data with $a_{ij} = b_{ij}$ for all $i = 1, \dots, m$, $j = 1, \dots, p$, it follows that the number of nontrivial intervals $q_i = 0$ and hence $n_i = 1$ for all $i = 1, \dots, m$. 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 $\boldsymbol{V}^* = (v_{j_1,j_2}^*)$, $j_1, j_2 = 1, \dots, p$,

$$\boldsymbol{V}^* = \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} \tag{2.18}$$

where $\boldsymbol{X}$ and $\boldsymbol{D}$ are as defined in (2.4) and (2.17), 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

$$\bar{X}_j^* = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w_k^i x_{kj}^i. \tag{2.19}$$

We can write (2.19) as

$$\bar{X}_j^* = \sum_{i=1}^{m}(\alpha_{ij}^a a_{ij} + \alpha_{ij}^b b_{ij}) \tag{2.20}$$

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 $t = a, b$,

$$\alpha_{ij}^t = \sum_{k=1}^{n_i} w_k^i = w_{ij}^t w_i \text{ whenever } x_{kj}^i = t_{ij} \tag{2.21}$$

It follows from (2.13) that

$$\alpha_{ij}^a + \alpha_{ij}^b = w_i. \tag{2.22}$$

Then, the variance v^*_{jj} of X_j can be written as

$$v^*_{jj} = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w^i_k (x^i_{kj} - \bar{X}^*_j)^2; \tag{2.23}$$

hence,

$$v^*_{jj} = \sum_{i=1}^{m} [\alpha^a_{ij}(a_{ij} - \bar{X}^*_j)^2 + \alpha^b_{ij}(b_{ij} - \bar{X}^*_j)^2]. \tag{2.24}$$

Likewise, the covariance $v^*_{j_1 j_2}$ between X_{j_1} and X_{j_2} can be written as

$$v^*_{j_1 j_2} = \sum_{i=1}^{m}\sum_{k=1}^{n_i} w^i_k (x^i_{kj_1} - \bar{X}^*_{j_1})(x^i_{kj_2} - \bar{X}^*_{j_2}). \tag{2.25}$$

We can show that

$$\begin{aligned} v^*_{j_1 j_2} &= \sum_{i=1}^{m} w_i (w^a_{ij_1} w^a_{ij_2} a_{ij_1} a_{ij_2} + w^a_{ij_1} w^b_{ij_2} a_{ij_1} b_{ij_2} + w^b_{ij_1} w^a_{ij_2} b_{ij_1} a_{ij_2} + w^b_{ij_1} w^b_{ij_2} b_{ij_1} b_{ij_2}) \\ &= \sum_{i=1}^{m} w_i x^0_{ij_1} x^0_{ij_2}, \end{aligned} \tag{2.26}$$

from (2.14). Hence, the variance-covariance matrix $\boldsymbol{V}^*$ based on the vertices is calculated.

Consider now the variance-covariance matrix based on the centers used in the $\boldsymbol{X}^c$ matrix, defined by

$$\boldsymbol{V}^c = (\boldsymbol{X}^c)^T \boldsymbol{D}^c \boldsymbol{X}^c \tag{2.27}$$

with elements $(v^c_{j_1 j_2})$, $j_1, j_2 = 1, \ldots, p$, and weight matrix $\boldsymbol{D}^c$. Then, the sample variance v^c_{jj} of the random variable X_j is

$$v^c_{jj} = \sum_{i=1}^{m} w_i (x^0_{ij} - \bar{X}^c_j)^2 \tag{2.28}$$

where the sample mean of the centers for X_j is

$$\bar{X}^c_j = \sum_{i=1}^{m} w_i x^0_{ij}. \tag{2.29}$$

For example, if uniformity within each interval holds and each observation is equally weighted, $x^0_{ij} = (a_{ij} + b_{ij})/2$, $w_i = 1/m$; and so (2.29) becomes $\bar{X}^c_j = \sum_i (a_{ij} + b_{ij})/(2m)$ as derived for intervals in Bertrand and Goupil (2000).

We can show that (2.28) becomes

$$v^c_{jj} = \sum_{i=1}^{m} w_i (w^a_{ij} a_{ij} + w^b_{ij} b_{ij})^2. \tag{2.30}$$

Likewise, the covariance $v^c_{j_1 j_2}$ between the variables X_{j_1} and X_{j_2} is

$$v^c_{j_1 j_2} = \sum_{i=1}^{m} w_i (x^0_{ij_1} - \bar{X}^c_{j_1})(x^0_{ij_2} - \bar{X}^c_{j_2}),$$

i.e.,

$$v^c_{j_1 j_2} = \sum_{i=1}^{m} w_i (w^a_{ij_1} a_{ij_1} + w^b_{ij_1} b_{ij_1})(w^a_{ij_2} a_{ij_2} + w^b_{ij_2} b_{ij_2}). \tag{2.31}$$

Hence, the variance-covariance matrix $\boldsymbol{V}^c$ based on the centers is obtained.

3 Vertices Principal Component Analysis

3.1 The Method

The data matrix $\boldsymbol{X}$ of (2.4) is considered to be the data matrix of n classical point observations on the random variables $(X_1, \ldots, X_p)$, with its associated weighted variance-covariance matrix being $\boldsymbol{V}^*$ of (2.18). 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., Jolliffe (1986), Johnson and Wichern (2002) for an applied presentation, and Anderson (1984) for a theoretical approach. Briefly, let $(\boldsymbol{e}_\nu, \lambda_\nu)$, with $\boldsymbol{e}_\nu = (e_{\nu 1}, \ldots, e_{\nu p})$, $\nu = 1, \ldots, p$, and with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, be the eigenvectors and eigenvalues of the matrix $\boldsymbol{V}^*$ suitably diagonalized. Then, the νth principal component $PC\nu$, $\nu = 1, \ldots, p$, satisfies

$$PC\nu = e_{\nu 1} X_1 + \cdots + e_{\nu p} X_p. \tag{3.1}$$

The total variance is $\sigma^2_n = \sum_{i=1}^{p} \lambda_i$ and the proportion of the total variance explained by $PC\nu$ is $\lambda_\nu / \sum_{\nu=1}^{p} \lambda_\nu$. Therefore, rather than focusing on the variations across all principal components, it becomes convenient to focus on the first $s \ll p$ principal components which typically together explain a large proportion of the total variation.

For the symbolic observation $\boldsymbol{\xi}_i$ represented by the n_i vertices in $\boldsymbol{X}_{\xi_i}$, the νth symbolic vertices principal component is obtained from

$$Y^*_{i\nu} = [y^a_{i\nu}, y^b_{i\nu}], \quad \nu = 1, \ldots, s \leq p, \tag{3.2}$$

where

$$y^a_{i\nu} = \min_{k \in L_i} \{y^i_{\nu k}\}, \quad y^b_{i\nu} = \max_{k \in L_i} \{y^i_{\nu k}\} \tag{3.3}$$

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 $y^i_{\nu k}$ is the value of the νth principal component for the row k in L_i. We can show that

$$y^a_{i\nu} = \sum_{j \in J^+}^{p} e_{\nu j}(a_{ij} - \bar{X}^*_j) + \sum_{j \in J^-}^{p} e_{\nu j}(b_{ij} - \bar{X}^*_j) \tag{3.4}$$

$$y^b_{i\nu} = \sum_{j \in J^-}^{p} e_{\nu j}(a_{ij} - \bar{X}^*_j) + \sum_{j \in J^+}^{p} e_{\nu j}(b_{ij} - \bar{X}^*_j) \tag{3.5}$$

where $J^+ = \{j | e_{\nu j} > 0\}$ and $J^- = \{j | e_{\nu j} < 0\}$.

A graphical representation of a set of s $(= 3)$ principal components obtained from a $p = 3$ dimensional observation is displayed in Figure 2. We observe the projection of the hypercube H_i onto the $PC1$ and $PC2$ plane (and also onto the $PC2$ and $PC3$ plane). 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 6 below.

The result of (3.2) and (3.3) can be verified as follows. Take any point $\tilde{\boldsymbol{x}}_i$ with $\tilde{x}_{ij} \in [a_{ij}, b_{ij}]$. Then, the νth principal component associated with this $\tilde{\boldsymbol{x}}_i$ is

$$\tilde{PC\nu} = \sum_{j=1}^{p} e_{\nu j}(\tilde{x}_{ij} - \bar{X}_j^*).$$

It follows that

$$\sum_{j=1}^{p} e_{\nu j}(\tilde{x}_{ij} - \bar{X}_j^*) \geq \sum_{j \in J^+}^{p} e_{\nu j}(a_{ij} - \bar{X}_j^*) + \sum_{j \in J^-}^{p} e_{\nu j}(b_{ij} - \bar{X}_j^*) \tag{3.6}$$

and

$$\sum_{j=1}^{p} e_{\nu j}(\tilde{x}_{ij} - \bar{X}_j^*) \leq \sum_{j \in J^-}^{p} e_{\nu j}(a_{ij} - \bar{X}_j^*) + \sum_{j \in J^+}^{p} e_{\nu j}(b_{ij} - \bar{X}_j^*) \tag{3.7}$$

However, by definition (3.3) and from (3.4), the right-hand side of (3.6) is

$$\min_{k \in L_i} \{y_{\nu k}^i\} = y_{i\nu}^a$$

and from (3.5), the right-hand side of (3.7) is

$$\max_{k \in L_i} \{y_{\nu k}^i\} = y_{i\nu}^b.$$

Hence, for all $\nu = 1, \ldots, p$,

$$\tilde{PC\nu} \in [y_{i\nu}^a, y_{i\nu}^b];$$

and so $Y_{i\nu}^*$ as in (3.2) and (3.3) holds for all $x_{ij} \in [a_{ij}, b_{ij}]$.

As for classical analyses, we can obtain a correlation measure between the νth principal component $PC\nu$ and the random variable X_j as

$$C_{j\nu} = Cor(X_j, PC\nu) = e_{\nu j}\sqrt{\lambda_\nu / \sigma_j^2} \tag{3.8}$$

where $e_{\nu j}$ is the component of the νth eigenvector $\boldsymbol{e}_\nu$ associated with X_j (see (3.1)) and where

$$\lambda_\nu = Var(PC\nu) \tag{3.9}$$

is the νth eigenvalue; and σ_j^2 is the variance of X_j. Note that when the variance-covariance matrix is standardized these σ_j^2 reduce to $\sigma_j^2 = 1$.

4 Centers Principal Component Analysis

4.1 The Method

An alternative method is one in which the values for each observation $\boldsymbol{\xi}_i$ are replaced by the barycenters $\boldsymbol{x}_i^0 = (x_{i1}^0, \ldots, x_{ip}^0)$ where

$$x_{ij}^0 = \sum_{k=1}^{n_i} (\frac{w_k^i}{w_i}) x_{kj}^i \tag{4.1}$$

where as before the observation $\boldsymbol{\xi}_i$ forms the hypercube H_i with vertices $k = 1, \ldots, n_i$, with weights w_k^i on the vertex $\boldsymbol{x}_k^i = (x_{k1}^i, \ldots, x_{kp}^i)$. It follows that, from (2.14),

$$x_{ij}^0 = w_{ij}^a a_{ij} + w_{ij}^b b_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, p. \tag{4.2}$$

Then, we construct the data matrix

$$\boldsymbol{X}^c = \begin{pmatrix} x_{11}^0 & \cdots & x_{1p}^0 \\ \vdots & & \vdots \\ x_{m1}^0 & \cdots & x_{mp}^0 \end{pmatrix}. \tag{4.3}$$

A special case is when $x_{ij}^0 \equiv x_{ij}^c$ of (2.5) and (2.6).

An advantage of the use of this $\boldsymbol{X}^c$ over the $\boldsymbol{X}$ used in the vertices method is that if the dimension p is particularly large, then use of the $n \times n$ $\boldsymbol{X}$ matrix is cumbersome. However, except for its entry in establishing the weights w_k^i used in (4.1) and the w_i in (4.2), much of the internal variation in the observations is lost. Thus, the total variation of the principal components is not observed and needs to be estimated from other descriptive measures. We compare the two methods more formally in Section 5.

The weight matrix is the $m \times m$ diagonal matrix

$$\boldsymbol{D}^c = diag(w_1, \ldots, w_m) \tag{4.4}$$

where the w_i take values which may incorporate the internal variations of the observations such as those of (2.7)-(2.11); see Section 2.2.

The centers principal components analysis is conducted by doing a classical analysis on the centered point observations of $\boldsymbol{X}^c$ after standardizing the variance-covariance matrix $\boldsymbol{V}^c$ of (2.28). Let $u_\nu = (u_{\nu 1}, \ldots, u_{\nu p})$, $\nu = 1, \ldots, p$, be the resulting νth centered eigenvector. Then, the νth centers principal component can be written as

$$PC\nu^c = \sum_{j=1}^{p} (x_j^0 - \bar{X}_j^c) u_{\nu j}. \tag{4.5}$$

In particular, let $\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})$ be any point contained in the hypercube H_i described by $\boldsymbol{\xi}_i$. Thus, we can calculate the νth centers principal component for this $\tilde{\boldsymbol{x}}_i$ from (4.5) as

$$PC\nu^c(\tilde{\boldsymbol{x}}_i) = \sum_{j=1}^{p} (\tilde{x}_{ij} - \bar{X}_j^c) u_{\nu j}.$$

Therefore, we can define the νth centers principal component as

$$Z_{i\nu} = [z^a_{i\nu}, z^b_{i\nu}], \quad \nu = 1, \ldots, s \leq p,$$

where

$$z^a_{i\nu} = \sum_{j=1}^{p} \min_{a_{ij} < \tilde{x}_{ij} < b_{ij}} \{(\tilde{x}_{ij} - \bar{X}^c_j) u_{\nu j}\} \tag{4.6}$$

and

$$z^b_{i\nu} = \sum_{j=1}^{p} \max_{a_{ij} < \tilde{x}_{ij} < b_{ij}} \{(\tilde{x}_{ij} - \bar{X}^c_j) u_{\nu j}\}. \tag{4.7}$$

It can be shown that these reduce to

$$z^a_{i\nu} = \sum_{j \in J^-_c}^{p} (b_{ij} - \bar{X}_j) u_{\nu j} + \sum_{j \in J^+_c}^{p} (a_{ij} - \bar{X}_j) u_{\nu j} \tag{4.8}$$

and

$$z^b_{i\nu} = \sum_{j \in J^-_c} (a_{ij} - \bar{X}_j) u_{\nu j} + \sum_{j \in J^+_c}^{p} (b_{ij} - \bar{X}_j) u_{\nu j} \tag{4.9}$$

where $J^-_c = \{j | u_{\nu j} < 0\}$ and $J^+_c = \{j | u_{\nu j} > 0\}$.

The graphical representation and the parameter interpretations for the centers principal components are analogous to their counterparts for the vertices principal components; see Section 3, and Figure 2.

5 Practical Considerations

5.1 Complexity and Information

Let us consider the computational complexity of the proposed approaches as it pertains to the efficiency of the calculation of the variance-covariance matrices, where we define complexity as the number of elementary operations needed to do this calculation. We recall that for the vertices method, each observation is represented by its n_i vertices; thus, 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. In contrast, since the centers method uses only the centroid ($\boldsymbol{x}^c$, or $\boldsymbol{x}^0$) for each observation, its order of complexity in calculating $\boldsymbol{V}^c$ is only $O(m)$. Once the $\boldsymbol{V}$, or $\boldsymbol{V}^c$, matrix is calculated, the computational effort is the same as for a classical principal component analysis based on p variables.

However, the centers method loses information in the data. For example, for $p = 1$ the two intervals [12, 16] and [2, 26] are quite different, but have the same central value $x^c = 14$ (if assume uniformity). The centers method uses the $x^c = 14$ value in both cases, whereas the vertices method by using the vertices captures the between and relative internal variations.

The key issue then is how we can retain the greater information inherent to the vertices method yet use the reduced computational complexity of the centers method. The answer is found in the mathematical results of the next subsection 5.2.

5.2 Comparison of the Vertices and Centers Methods Statistics

We can show that, from (2.14) and (2.19)-(2.21),

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

i.e.,

$$\bar{X}_j^c = \bar{X}_j^*; \tag{5.1}$$

that is, the weighted means are the same for both methods, for all $j = 1, \ldots, p$.

However, the variance-covariance matrices differ for the two methods. From (2.23) and (2.28), the difference between the two variances is

$$v_{jj}^* - v_{jj}^c = \sum_{i=1}^{m} [\alpha_{ij}^a (a_{ij} - \bar{X}_j^*)^2 + \alpha_{ij}^b (b_{ij} - \bar{X}_j^*)^2] - \sum_{i=1}^{m} w_i (w_{ij}^a a_{ij} + w_{ij}^b b_{ij})^2. \tag{5.2}$$

We can show that (5.2) becomes, for $j = 1, \ldots, p$,

$$v_{jj}^* - v_{jj}^c = \sum_{i=1}^{m} w_i w_{ij}^a w_{ij}^b (b_{ij} - a_{ij})^2 = e_{jj}, \text{ say;} \tag{5.3}$$

i.e.,

$$v_{jj}^* = v_{jj}^c + e_{jj}, \quad j = 1, \ldots, p. \tag{5.4}$$

The factor e_{jj} represents the loss of precision or variation that pertains when the interval $[a_j, b_j]$ is replaced by its centroid x_j^0. When the differences $(b_j - a_j)$ are small, the two methods would clearly give similar results.

Consider now the covariances between X_{j_1} and X_{j_2} for the two methods. Expanding (2.31), we have

$$\begin{aligned} v_{j_1 j_2}^c &= \sum_{i=1}^{m} w_i (w_{ij_1}^a w_{ij_2}^a a_{ij_1} a_{ij_2} + w_{ij_1}^a w_{ij_2}^b a_{ij_1} b_{ij_2} \\ &\quad + w_{ij_1}^b w_{ij_2}^a b_{ij_1} a_{ij_2} + w_{ij_1}^b w_{ij_2}^b b_{ij_1} b_{ij_2}) \\ &= v_{j_1 j_2}^* \end{aligned} \tag{5.5}$$

from (2.25). That is, for $j_1 \neq j_2$, the covariances are the same for both methods.

Hence, from (5.1)-(5.5), we have that the vertices variance-covariance matrix $\boldsymbol{V}^*$ and the centers variance-covariance matrix $\boldsymbol{V}^c$ satisfy the relationship

$$\boldsymbol{V}^* = \boldsymbol{V}^c + \boldsymbol{E} \tag{5.6}$$

where $\boldsymbol{E}$ is a $p \times p$ diagonal matrix with diagonal elements e_{jj} given by (5.3). Thus, we see that the variance of the vertices data has two components, one representing the interval variation within each observation (i.e., the interval of length (b-a)) and the other representing variation between (the centroids of the) observations.

Classical Data:

When the data are all classical observations, we have from (5.3), $e_{jj} = 0$, for all j. In this case, the two methods are equivalent throughout and the classical principal component analysis becomes a special case.

5.3 A Refinement

The relationship (5.6) allows for the calculation of the vertices variance-covariance matrix $\boldsymbol{V}^*$ by calculating the centers 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 (2.18).

A vertices principal component analysis is then performed with a variance-covariance matrix V^* obtained with a complexity of $O(m)$. Interval-valued principal components are also given by (3.4) and (3.5) with a complexity of $O(m)$ instead of obtaining them from (3.3) directly (with its complexity $O(m2^p)$). 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.

6 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 νth principal component axis, viz.,

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

where $d(X_i, G)$ is the Euclidean distance between the observation X_i and G is the centroid of all data $X_i, i = 1, \cdots, n$ values. Large values of $cos(X_i, PC\nu)$ mean that the position of X_i is near to its projected value on the $PC\nu$ 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 $cos(X_i, PC\nu)$ suggest care is necessary when interpreting results relative to that X_i and $PC\nu$.

A second quantity useful for interpretation purposes in a classical analysis is the contribution of each observation X_i to the inertia, viz.,

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

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 $PC\nu$ by an observation $\boldsymbol{\xi}_i$ represented here by its observed hypercube H_i can be measured by

$$C^1_{i\nu} = Ctr(H_i, PC\nu) = w_i \sum_{k=1}^{n_i} \frac{w^i_k (y^i_{\nu k})^2}{[d(\boldsymbol{x}^i_k, \boldsymbol{G})]^2}. \tag{6.1}$$

where $y^i_{\nu k}$ is the νth principal component for the vertex k of H_i (see (3.3)), w^i_k is the weight of that vertex (see Section 2.2), and where $d(\boldsymbol{x}^i_k, \boldsymbol{G})$ is the Euclidean distance between the vertex $\boldsymbol{x}^i_k$ 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

$$C^2_{i\nu} = Ctr(H_i, PC\nu) = \frac{\sum_{k=1}^{n_i} w^i_k (y^i_{\nu k})^2}{\sum_{k=1}^{n_i} w^i_k [d(\boldsymbol{x}^i_k, \boldsymbol{G})]^2} \tag{6.2}$$

The first function $C^1_{i\nu}$ identifies the average squared cosines of the angles between these vertices and the axis of the νth principal component. The second function $C^2_{i\nu}$ 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 the relative contributions $C^2_{i\nu}$ of (6.2) are smaller than the $C^1_{i\nu}$ of (6.1).

The absolute contribution of a single observation through the vertices of H_i to this variance λ_ν is measured by the inertia

$$I_{i\nu} = \text{Inertia}(H_i, PC\nu) = [\sum_{k=1}^{n_i} w^i_k (y^i_{\nu k})^2]/\lambda_\nu \tag{6.3}$$

and the contribution of this observation to the total variance is

$$I_i = \text{Inertia}(H_i) = \{\sum_{k=1}^{n_i} w^i_k [d(\boldsymbol{x}^i_k, \boldsymbol{G})]^2\}/I_T, \tag{6.4}$$

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

$$\sum_{i=1}^{m} I_{i\nu} = \lambda_\nu, \quad \sum_{i=1}^{m} I_i = I_T. \tag{6.5}$$

In a different direction, an alternative visual aid in interpreting the results is that whereby only those vertices whose contribution to the principal component $PC\nu$ exceed some prespecified value α, be used in the equations (3.3). That is, we set

$$Y^*_{i\nu}(\alpha) = [y^a_{i\nu}(\alpha), y^b_{i\nu}(\alpha)]$$

where

$$y^a_{i\nu}(\alpha) = \min_{k \in L_i} \{y^i_{\nu k} | Ctr(\boldsymbol{x}^i_k, PC\nu) \geq \alpha\}, \quad y^b_{i\nu}(\alpha) = \max_{k \in L_i} \{y^i_{\nu k} | Ctr(\boldsymbol{x}^i_k, PC\nu) \geq \alpha\}, \tag{6.6}$$

where

$$Ctr(\boldsymbol{x}_k^i, PC\nu) = \frac{(y_{\nu k}^i)^2}{[d(\boldsymbol{x}_k^i, \boldsymbol{G})]^2} \tag{6.7}$$

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

Consider the vertices method (similar arguments hold for the centers method). When $\alpha = 0$, the symbolic principal component interval obtained from (3.3) 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 (6.7). Thus, to take an extreme case, one vertex $k = k'$ may have a value of $y_{\upsilon k'}^i = 1.0$ (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 (3.3) gives $PC\nu = [1.0,\ 11.0]$. Suppose however the relative contribution, from (6.7), 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 (6.6), 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 (3.3) 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 hold.

An alternative to the criterion of (6.6) is to replace $Ctr(\boldsymbol{x}_k^i, PC\nu)$ by

$$Ctr(\boldsymbol{x}_k^i, PC\nu_1, PC\nu_2) = Ctr(\boldsymbol{x}_k^i, PC\nu_1) + Ctr(\boldsymbol{x}_k^i, PC\nu_2). \tag{6.8}$$

In this case, vertices that make larger contributions in either of the two principal components $PC\nu_1$ and $PC\nu_2$ 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 Figure 3. The principal component envelope (obtained from applying (3.4)-(3.5)) 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 (6.7). 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 (6.6), the vertex contributing 0.05, is omitted, with the resulting principal component envelope being that shown in Figure 4. The observation represented by the hypercube H_2 has six vertices (see Figure 3) including a vertex whose contribution is 0.01. Application of (6.6) 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 $Ctr(\boldsymbol{x}_k^i,\ PC\nu_1) < \alpha$, 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. 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.

$$\bar{y}_{i\nu_1} = \frac{1}{n_i}\sum_{k=1}^{n_i} y^i_{\nu_1 k}. \tag{6.9}$$

7 Constrained Observations

It can be that certain constraints are imposed on observational values. Some constraints, e.g., age ≥ 50, are common to both classical and symbolic data. However, depending on the nature of (any) aggregation of an original dataset that produced a symbolic dataset, we may need to impose rules or constraints to maintain the integrity of the original values themselves. For example, suppose a dataset contains values for age and number of children for individuals living in various cities, and suppose that no person under the age of 12 (say) had any children. Suppose now the data are aggregated by city. Since clearly there can be individuals over 12 with no children, any such aggregation can produce a symbolic value of, e.g, age = $[5, 70]$ and number of children = $[0, 1, 2, \ldots]$. In this case, the associated hypercube includes the point, age = 5 and number of children = 2; and so on. A logical dependency rule that "If age <12, then number of children = 0" preserves the integrity of the original values. These constraints in effect produce "holes" in the hypercubes; see Figure 5. For another example, suppose two variables X_1= number of at-bats and X_2= number of hits, over a season of baseball. Suppose the aggregation over all players in a given team produced the $p = 2$ dimensional hypercube H=([80, 400], [30, 180]), a rectangle. However, since for any one player $x_2 \leq x_1$, observations within the triangle with vertices (80, 80), (80, 180), and (180, 180) are not possible. Thus, this triangle represents a "hole" in the hypercube H, leaving a 5-sided hypercube in $\Re^2$ as the valid dataset or "constrained hypercube"; see Billard and Diday (2006).

The foregoing construction of the data matrix $\boldsymbol{X}$ and $\boldsymbol{X}^c$ and their weights of Section 2.1 and 2.2 are adjusted for constraints as follows. Suppose the effect of a constraint on the validity of the hypercube H_i associated with an observation $\boldsymbol{\xi}_i$ is such that a subhypercube (or constrained hypercube) C^i is not valid (i.e., C^i is a "hole" in H_i), and suppose C^i can be expressed by

$$C^i : ([ca_{ij_1}, cb_{ij_1}]\wedge, \ldots, \wedge[ca_{ij_r}, cb_{ij_r}],\ [ca_{ij_k}, cb_{ij_k}] \subseteq [a_{ij_k}, b_{ij_k}]\ j_k \in \{1, \ldots, p\}), \tag{7.1}$$

where $[ca_{ij}, cb_{ij}]$ are the constrained interval values for the variable X_j on observation $\boldsymbol{\xi}_i$, and where clearly for all $j_k \in \{1, \ldots, p\}$, $(x_{ij_1}, \ldots, x_{ij_r}) \in C^i$ if and only if

$$x_{ij_k} \in\ [ca_{ij_k},\ \ cb_{ij_k}]. \tag{7.2}$$

Note that there can be more than one constrained interval for any one variable j. These constrained hypercubes necessarily reside inside H_i. There can be several constrained regions within H_i, viz., $C^i_1, \ldots, C^i_{r_i}$. The example in Figure 5 has $r_i = 2$. Here, $n^1_i = 8$

and $n_i^2 = 4$ vertices, respectively, in C_1^i and C_2^i. For clarity of presentation, we assume these constrained hypercubes are disjoint. As for the hypercube H_i itself, these constrained hypercubes can be points, lines, and so on, with n_i^r vertices in the constrained hypercube C_r^i, $r = 1, \ldots, r_i$.

One impact of these constraints is on the values for the weights w_i of the observation $\boldsymbol{\xi}_i$ and for weights w_k^i of the k vertex of H_i. One proposed set of new weights is found by extending the approach used in Section 2.2 as follows.

The volume of the constrained hypercube C_r^i is, for $r = 1, \ldots, r_i$, $i = 1, \ldots, m$,

$$V(C_r^i) = \prod_{\substack{j \in E_r^i \\ ca_{ij}^r \neq cb_{ij}^r}} (cb_{ij}^r - cb_{ij}^r), \tag{7.3}$$

where E_r^i is the set $j_k \in \{1, \ldots, p\}$ of variables in the constrained hypercube C_r^i, and where $[ca_{ij}^r, cb_{ij}^r]$ are as defined in (7.1) for C_r^i with the r superscript added.

Then, the new weight for $\boldsymbol{\xi}_i$ is

$$w_i^* = w_i[1 - V_i^{-1} \sum_{r=1}^{r_i} V(C_r^i)], \quad i = 1, \ldots, m, \tag{7.4}$$

where the volume V_i was defined in (2.10). When, for a given i, there are no holes in the corresponding H_i, i.e., when the observation $\boldsymbol{\xi}_i$ is unconstrained, then from (7.4), $w_i^* = w_i$.

Finally, new weights for the vertices are given as

$$(w_k^i)^* = w_i^* d(\boldsymbol{x}_k^i) / \sum_{k=1}^{n_i} d(\boldsymbol{x}_k^i) \tag{7.5}$$

where

$$d(\boldsymbol{x}_k^i) = \frac{1}{r_i} \sum_{r=1}^{r_i} D(\boldsymbol{x}_k^i, C_r^i) \tag{7.6}$$

with

$$D(\boldsymbol{x}_k^i, C_r^i) = \frac{1}{n_i^r} \sum_{l=1}^{n_i^r} d(\boldsymbol{x}_k^i, \boldsymbol{x}_l^{C_r^i}). \tag{7.7}$$

The distance $d(\boldsymbol{x}_k^i, \boldsymbol{x}_l^{C_r^i})$ in (7.7) is the Euclidean distance measured from the kth vertex in H_i (denoted by $\boldsymbol{x}_k^i$) and the collection of vertices in C_r^i (denoted by $\boldsymbol{x}_l^{C_r^i}$). Thus, $D(\boldsymbol{x}_k^i, C_r^i)$ is the (collective) average such distance averaged over all vertices in C_r^i. Therefore, the distance $d(\boldsymbol{x}_k^i)$ in (7.6) is the average of the collective average distances for each constrained hypercube C_r^i averaged over all the constrained C_r^i's for the observation $\boldsymbol{\xi}_i$.

It can be verified that, as required,

$$\sum_{k=1}^{n_i} (w_k^i)^* = w_i^*, \qquad \sum_{i=1}^{m} w_i^* = 1. \tag{7.8}$$

The vertices and/or centers methods then proceed as in Section 3 and/or Section 4 but with these weights being used instead of those obtained when no constraints exist.

8 Face Recognition Application

8.1 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 (1973) study of matching pictorial structures, followed by Baron (1981), among others. Following a brief review by Samal and Iyengar (1992), in an excellent and extensive review, Chellappa et al. (1995) 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. (2003) 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 (1991), Craw and Cameron (1996) and Moon and Phillips (2001), discriminant analysis by Eternad and Chellappa (1997), probabilistic eigenfaces developed by Moghaddam and Pentland (1997), and nearest line features considered by Li and Chellappa (2002) and Li and Lu (1999). Studies such as those by Kass et al. (1987), Turk (1991), Craw et al. (1992) and Staib and Duncan (1992) helped identify those facial features that should be included in any discrimination research. Zhao et al. (2003) conclude that while progress has been valuable, much more remains to be done especially when databases are large.

Our analysis will focus on a dataset from an investigation by Leroy et al. (1996) 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, ...). 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 Figure 6), 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 each extracted boundaries are localized, and then a distance is measured between a specific pair of points as represented by these random variables, in Figure 6. 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.

Table 1 - Faces Dataset (Distances AD,. . . ,GH as in Figure 6, see text)

Subject	X_1 = AD	X_2 = BC	X_3 = AH	X_4 = DH	X_5 = EH	X_6 = 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]

Before carrying out the analysis, let us first make the following comment that pertains for aggregated data such as in the faces data. As described, there are 27 interval-valued observations. Suppose each observation drew from a sequence of 1000 images. This gives a total of 27000 classical point observations in $\Re^6$. An underlying assumption of the standard classical analysis is that all 27000 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 27000 observations. The resulting principal component analysis will look for axes which maximize the variability across all 27000 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.

8.2 Vertices Principal Components Analysis

We first apply the vertices principal component method to the data of Table 1. Observations and their vertices were given equal weights ((2.8) and (2.12)). Values of the first three vertices principal components obtained through the application of (3.4)-(3.5) for each observation are displayed in Table 2.

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]

The plots of these along the first principal component ($PC1$) and second principal component ($PC2$) axes are shown in Figure 7. 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. The faces {INC, FRA} might be one class; the faces {HUS, PHI, JPL} suggest themselves as another class, as do the faces {ISA, ROM}; and the faces {LOT, KHA} would be a fourth class.

By restricting the calculation of the principal components to those vertices which have a contribution α or more, i.e., by using (6.6), we can obtain a clearer picture of the class groupings. The relative contribution $Ctr(\boldsymbol{x}_k^i, PC\nu)$, $\nu = 1, 2, k = 1, \ldots, n_i$ are calculated from (6.7) 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 $Ctr(\boldsymbol{x}_k^i, PC1) > 0.2$. Therefore, all (of the $2^6 = 64$ total) vertices enter into the application of (6.6) to give us

$$PC1(\alpha = 0.2) = [-3.662, -1.354].$$

However, for $\nu = 2$, only 8 of the 64 vertices satisfy the relation $Ctr(\boldsymbol{x}_k^i, PC2) > 0.2$. 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 (6.6). Hence, we obtain the second vertices principal component as

$$PC2(\alpha = 0.2) = [-2.051, -1.644].$$

Table 3 - Vertices Contributions to $PC\nu = 1, 2$ ($i = 8 \equiv INC2$): Faces

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 provides the complete set of vertices principal components obtained from (6.6) when $\alpha = 0.2$; these are plotted in Figure 8. Also, given 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 vertex contributes more than 0.2 to the second principal component. In this case, $\bar{y}_{7,2} = 0.28$ from (6.9). This is reflected as a line (instead of a rectangle) parallel to the first principal component axis in Figure 8. Notice that the face JPL1 also assumes a linear form (parallel to the second principal component axis). In this case, however, this arises from (6.6) where now only one vertex contributes more than $\alpha = 0.2$ to the first principal component.

By comparing the principal components of Figure 7 (which corresponds to $\alpha = 0.0$) and of Figure 8 (where $\alpha = 0.2$), the greater clarity of the classes that emerges 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 be divided into two, $\{PHI, JPL\}$ and $\{HUS\}$.

Table 4 - Vertices Principal Components, $\nu = 1, 2$, $\alpha = 0.2$: Faces

	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

Table 5 gives all the eigenvalues λ_ν, $\nu = 1, \ldots, 6$ (from (3.9), 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 5 - Vertices PC Inertia: Faces

$PC\nu$	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

The correlations $C_{j\nu}$ between the variable X_j and the νth principal component $PC\nu$ were calculated from (3.8) 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 6 - Vertices Method, Correlations $C_{j\nu}$ between X_j and $PC\nu$: Faces

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

Based on 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 $Ctr(H_i, PC_\nu)$ between the full observation $\boldsymbol{\xi}_i$ and the νth principal component. These values, calculated from (6.2), 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. The characteristics of the other groupings can likewise be identified.

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 Figure 7, that the ISA faces essentially form their own cluster.

Table 7 - Vertices Method, Relative Contributions to Vertices $PC\nu$, $\nu = 1, 2, 3$: Faces

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

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 $PC\nu$, $\nu = 1, 2, 3$, for each observation, obtained from (6.3). Then, in the right-hand column, we give this contribution I_i of each observation to the total variance, calculated from (6.4). Thus, e.g., 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.

Table 8 - Vertices Method, Absolute Contributions of Subject to $PC\nu$ and Inertia: Faces

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

8.3 Centers Principal Components Analysis

The centers method of Section 4 when applied to the data of Table 1 produced the principal components shown in Table 9 for $\nu = 1, 2, 3$, obtained from (4.6)-(4.7), or (4.8)-(4.9). The plots against the axes $PC1$ and $PC2$ are shown in Figure 9. Again, it is immediately apparent that a coherence between the three sequences for each person exists. The same initial groups obtained by the vertices method emerge. However, this time when looking at the plots on the $PC2$ and $PC3$ axes, it is the group $\{ROM, ISA\}$ which splits into two groups $\{ROM\}$ and $\{ISA\}$.

Table 9 - Centers Principal Components, $\nu = 1, 2, 3$: Faces

	$PC1$	$PC2$	$PC3$
FRA1	[-2.969, -1.747]	[0.236, 1.722]	[-0.376, 1.033]
FRA2	[-2.729, -1.111]	[-0.197, 1.789]	[-0.372, 1.067]
FRA3	[-3.286, -0.856]	[-0.532, 2.077]	[-1.032, 1.301]
HUS1	[-0.374, 1.162]	[0.376, 2.284]	[0.772, 2.419]
HUS2	[-0.583, 1.482]	[0.592, 2.975]	[0.355, 2.592]
HUS3	[-0.403, 1.506]	[0.461, 2.822]	[0.384, 2.418]
INC1	[-4.065, -2.381]	[-0.902, 1.289]	[-0.905, 0.744]
INC2	[-3.909, -1.213]	[-2.509, 0.933]	[-0.954, 2.022]
INC3	[-4.332, -1.837]	[-1.495, 1.469]	[-1.115, 1.615]
ISA1	[0.881, 2.239]	[-2.063, -0.477]	[-0.603, 0.708]
ISA2	[0.388, 2.038]	[-1.933, -0.073]	[-0.688, 0.868]
ISA3	[0.444, 2.355]	[-2.088, -0.112]	[-0.618, 1.388]
JPL1	[-0.567, 0.888]	[0.667, 2.379]	[-2.068, -0.536]
JPL2	[-0.593, 1.167]	[0.575, 2.764]	[-2.101, -0.145]
JPL3	[-0.782, 0.916]	[0.593, 2.645]	[-1.805, 0.206]
KHA1	[-1.097, 0.641]	[-3.507, -1.653]	[-1.280, 0.368]
KHA2	[-1.429, 0.386]	[-3.645, -1.505]	[-0.993, 0.743]
KHA3	[-1.730, 0.468]	[-3.393, -0.817]	[-1.346, 0.714]
LOT1	[-0.794, 0.742]	[-2.864, -0.977]	[-0.874, 0.707]
LOT2	[-0.773, 0.472]	[-2.205, -0.687]	[-0.879, 0.376]
LOT3	[-0.928, 0.408]	[-2.435, -0.778]	[-0.848, 0.586]
PHI1	[0.114, 1.574]	[0.722, 2.178]	[-1.582, -0.030]
PHI2	[-0.270, 1.740]	[0.454, 2.689]	[-2.017, 0.220]
PHI3	[-0.450, 1.440]	[0.356, 2.671]	[-2.258, -0.168]
ROM1	[2.407, 3.838]	[-1.270, 0.436]	[-0.549, 0.902]
ROM2	[1.961, 4.041]	[-1.394, 1.200]	[-0.899, 1.491]
ROM3	[1.529, 3.978]	[-1.436, 1.585]	[-0.874, 1.918]

Table 10 - Relative Contributions to Centers $PC\nu = 1, 2, 3$: Faces

Subject	PC1	PC2	PC3	Subject	PC1	PC2	PC3
FRA1	0.74	0.13	0.01	KHA1	0.01	0.89	0.03
FRA2	0.71	0.12	0.02	KHA2	0.04	0.93	0.00
FRA3	0.83	0.12	0.00	KHA3	0.08	0.84	0.02
HUS1	0.03	0.37	0.54	LOT1	0.00	0.81	0.00
HUS2	0.04	0.56	0.38	LOT2	0.01	0.62	0.02
HUS3	0.06	0.53	0.38	LOT3	0.02	0.60	0.00
INC1	0.95	0.00	0.00	PHI1	0.18	0.52	0.16
INC2	0.75	0.07	0.03	PHI2	0.13	0.59	0.19
INC3	0.94	0.00	0.01	PHI3	0.06	0.51	0.33
ISA1	0.55	0.37	0.00	ROM1	0.89	0.02	0.00
ISA2	0.46	0.31	0.00	ROM2	0.94	0.00	0.01
ISA3	0.52	0.32	0.04	ROM3	0.88	0.00	0.03
JPL1	0.01	0.52	0.38				
JPL2	0.02	0.56	0.26				
JPL3	0.01	0.67	0.16				

We can calculate the relative contribution of each hypercube to each principal component through (6.2). The resulting values are given in Table 10. From these we observe that the three INC faces along with those of FRA and ROM all contribute strongly to the first principal component $PC1$, while those for KHA contribute most to the second principal component, followed by the faces of LOT and PHI and then HUS. These contributions and their conclusions are comparable to those obtained using the vertices principal component method.

The eigenvalues, percents of total variation and cumulative variation percents, for each principal component, are displayed in Table 11. The first two principal components account for 81% of the total variation (and can be contrasted with the 73% for the vertices method).

Table 11 - Centers *PC* Inertia: Faces

PC	Eigenvalue	% Inertia	Cumulative Inertia
$PC1$	2.788	46.5	46.5
$PC2$	2.044	34.1	80.6
$PC3$	0.547	9.1	89.7
$PC4$	0.324	5.4	95.1
$PC5$	0.234	3.9	99.9
$PC6$	0.062	1.0	100

Table 12 - Centers Method, Correlations $C_{j\nu}$ between X_j and PC_ν: Faces

X_j	$PC1$	$PC2$	$PC3$
AD	0.640	0.648	0.174
BC	0.496	0.736	-0.140
AH	0.862	-0.164	-0.418
DH	0.910	0.130	0.205
EH	-0.559	0.655	-0.464
GH	-0.500	0.780	0.255

Table 12 shows the correlations between the principal components for $\nu = 1, 2, 3$ and each of the X_j, $j = 1, \ldots, 6$, variables. Again, we identify the overall size variables $X_4 = DH$ and $X_3 = AH$ as the two main characterizing variables relative to the first principal component with correlations 0.910 and 0.862, respectively, plus that of the eye-span $X_1 = AD$ with a correlation of 0.640 as a contributing identifier. Relative to the second principal component, we again have the other three variables X_6 and X_2 (in that order) as the major identifying variables, with $X_5 = EH$ also part of the descriptor. However, unlike the vertices analysis, this time it is not unreasonable to add $X_1 = AD$ along with X_5 as additional descriptors.

Further analyses such as those conducted for the vertices method can also be performed. There are some internal differences; e.g., the relative contributions $Ctr(H_i, PC_\nu)$ of the subjects to the principal components are higher for the face $INC(I_{i1} = 0.91, 0.75, 0.94)$ than those obtained by the vertices method and some are lower such as $JPL(I_{i1} = 0.01, 0.02, 0.00)$. See Chouakria (1998) for details.

However, as before, the analyses' results allow the investigator to isolate those characteristics (variables) which help identify classes of faces.

8.4 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 analysis is 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 $m \times p$ $(= 27 \times 6)$ classical dataset. A plot of the first and second principal components for these data is shown in Figure 10. One limiting factor of this surrogate is that it is impossible to retain any measure of the internal variation; e.g., 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 11 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 (2006) in their analysis of fuzzy data. Thus, 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 Figure 12.

These three surrogate analyses are compared through Figures 10-12. While both Figure 10 and Figure 11 retain the coherences for the sets of the same three faces already observed for the symbolic analysis, the range surrogate in general loses that coherence (though it can be observed 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 Figure 10 for the midpoints with Figure 12 when ranges are also used along with the midpoint variable. It is also clear from Figure 12 that clusters are mixtures of faces, e.g., faces {INC1, INC3, FRA3} could be one cluster.

8.5 Comparison of Symbolic and Surrogate Analyses

For comparison purposes, consider the vertices symbolic principal component analysis and the midpoint surrogate analysis; similar conclusions pertain when the centers symbolic and 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 symbolic results reflect all the variations between the observations including the internal variations, whereas the classical results do not. The classical values plotted in Figures 10 are points in space, while the symbolic values are hypercubes, here rectangles for the $s = 2$ principal components plotted in Figure 7. These rectangles have smaller (or larger) dimensions whenever the original data are smaller (or larger) intervals. [This point will be discussed further in Section 9.] 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 the 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 Figure 10 and Figure 11).

For these data, the two analyses produce slightly different groups of faces (though arguments can be made for the same groupings - but see the bats analysis below where differences are more pronounced). 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 Figure 7, 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 Figure 8), where it is more obvious that the ISA and ROM faces belong to the same group. Notice too that, from Figure 8, 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 Figure 8 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.

9 Bats Dataset

9.1 The Data

The bats dataset displayed in Table 13 is an example of naturally occurring interval-valued data. There are four random variables, X_1 = head size, X_2 = tail length, X_3 = height, X_4 = forearm length; and there are $m = 21$ species (PIPC, ... , MGES, as shown; the species identifier is an abbreviation of the longer biological latin descriptor, e.g., 'BARB' is the species *Barbastella barbastellus*). A scientific question relates to whether or not certain species are alike. Since the data are naturally intervals, a symbolic principal component is required. We limit our symbolic analysis to the vertices method, with equal weights ((2.8) and 2.12)). Taking the interval midpoints as classical surrogates in a standard analysis will also be implemented, and shown to be quite inadequate in explaining the inherent variations in the data.

Table 13 - Bats Species Dataset

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]

9.2 Vertices Principal Components

By applying the vertices principal component analysis of Section 3, we obtain the first and second principal components from (3.4) - (3.5), as shown in Table 14 for each observation. These are plotted in Figure 13. Also shown in Table 14 are the principal components obtained from (6.6), for $\alpha = 0.4$, and from (3.4) - (3.5), as well as the number of vertices (from a possible $2^4 = 16$ vertices) which contributed a level of α or more to the principal components. These latter principal components are plotted in Figure 14. Seven groups emerge. The first group $G_1 = \{$NOCT, MGES, GMUR$\}$ is characterized by its large head and forearm measurements. The species in group $G_2 = \{$MSCH, SCOM$\}$ are characterized by their large tail; their opposite is group $G_3 = \{$MDAUB, PRH$\}$ with small tails. The set $G_4 = \{$PIPC, MOUS, PIPS$\}$ has a small head size and small forearm lengths. These contrast with $G_5 = \{$BARB, OGRIS$\}$ who are characterized by having small heights (as do the species in G_4) but whose other variables have larger measurements (than do those in G_4). The species $G_6 = \{$FCHEV$\}$ clearly separates itself out (by virtue of its being the largest species on height). Finally, group $G_7 = \{$PIPN, MNAT, MDEC, MGP, OCOM, MBEC, SBOR, SBIC$\}$ is clustered in the middle of these plots and is characterized by medium sized heads, forearms, tail, and height. When $\alpha = 0.4$, it becomes clear that G_1 and G_2 are really two distinct groups, rather than perhaps one as suggested from the plots with $\alpha = 0$, and also that these two groups together are quite distinct from the core middle grouping represented by the groups G_3, G_4 and G_7. The separation is even more apparent when $\alpha = 0.5$ (see Figure 15). Also, from Figure 15, it is more evident that the group G_3 is a separate group from the central core of species. It also suggests that the species $\{$MGP$\}$ is a separate group characterized by there being less variability on sizes within the species itself. Thus, by exploiting the strength of the alpha term, greater clarity can be attained as to the most appropriate clusters.

Table 14 - Vertices Principal Components, $\nu = 1, 2$, $\alpha = 0, 0.4$: Bats

		$\alpha = 0$		$\alpha = 0.4$			
	Species	$PC1$	$PC2$	$PC1$	$PC2$	# of Vertices	
1	PIPC	[-3.637, -1.612]	[-0.635, 1.032]	[-3.637, -1.612]	0.032	16	0
2	PRH	[-2.371, -0.711]	[-2.037, -0.511]	[-2.371, -1.362]	[-2.037, -1.284]	10	7
3	MOUS	[-2.307, -0.855]	[-0.633, 0.701]	[-2.307, -0.855]	0.077	16	0
4	PIPS	[-2.199, -1.139]	[-0.017, 0.881]	[-2.199, -1.139]	0.075	16	0
5	PIPN	[-1.942, -0.807]	[-0.216, 1.041]	[-1.942, -0.807]	[0.917, 0.917]	16	1
6	MDAUB	[-1.912, 0.025]	[-1.477, 0.344]	[-1.912, -0.626]	[-1.477, -0.627]	8	5
7	MNAT	[-1.603, -0.204]	[-0.809, 0.595]	[-1.603, -0.557]	[-0.731, -.447]	13	2
8	MDEC	[-1.140, -0.356]	[-0.243, 0.373]	[-1.140, -0.589]	0.057	12	0
9	MGP	[-0.734, 0.451]	[-1.352, -0.384]	[-0.734, -0.595]	[-1.352, -0.508]	2	14
10	OCOM	[-1.451, 0.848]	[-1.107, 0.982]	[-1.451, 0.848]	[-1.107, 0.982]	5	4
11	MBEC	[-0.958, 0.513]	[-1.168, 0.376]	[-0.958, 0.513]	[-1.168, 0.376]	5	7
12	SBOR	[-0.810, 0.575]	[-0.764, 0.686]	[-0.810, -0.516]	[-0.764, 0.686]	3	6
13	BARB	[-1.599, 0.349]	[0.246, 2.133]	[-1.599, -0.811]	[1.334, 2.133]	4	8
14	OGRIS	[-1.052, 0.321]	[0.185, 1.694]	[-1.052, -0.382]	[0.243, 1.694]	5	11
15	SBIC	[-0.678, 0.916]	[-0.444, 0.789]	[-0.678, 0.916]	[-0.444, 0.789]	3	3
16	FCHEV	[0.155, 2.536]	[-2.400, -0.366]	[0.759, 2.536]	[-2.400, -1.453]	8	7
17	MSCH	[0.418, 1.770]	[0.130, 1.450]	[0.735, 1.770]	[1.373, 1.450]	10	2
18	SCOM	[0.765, 3.112]	[-0.564, 1.502]	[0.764, 3.112]	[1.323, 1.323]	16	1
19	NOCT	[0.851, 3.226]	[-0.890, 1.505]	[0.851, 3.226]	0.115	15	0
20	GMUR	[2.040, 4.826]	[-1.696, 0.806]	[2.040, 4.826]	0.052	16	0
21	MGES	[2.581, 3.802]	[-0.544, 0.806]	[2.581, 3.802]	0.021	16	0

Again, the vertices principal components reflect the relative sizes of the original interval-valued data. Figure 16 isolates the first two principal components for the four species

BARB, OGRIS, GMUR, and MGES, for $\alpha = 0$. Clearly, the species OGRIS has a principal component hypercube that is smaller in size than is that for the species BARB. This reflects the shorter interval lengths for observed values for OGRIS over those for BARB. Indeed, these data are such that the actual data hypercube for OGRIS is almost (but not entirely) contained within that for BARB. The same reflections hold for the MGES and GMUR species.

Table 15 - Vertices *PC* Inertia: Bats

PC_ν	Eigenvalue λ_ν	% Inertia	Cumulative Inertia
$PC1$	2.708	67.7	67.7
$PC2$	0.687	17.2	84.9
$PC3$	0.392	9.8	94.7
$PC4$	0.213	5.3	100

The variation explained by each principal component along with the cummulative percentage of the total variation and the respective eigenvalues λ_ν, $\nu = 1, ..., 4$, are shown in Table 15. We see that 67.7% of the total variation is explained by the first vertices principal component and 84.9% is explained by the first and second principal components. Table 16 provides the correlations $C_{\jmath v}$ between the first three vertices principal components $PC\nu$, $\nu = 1,\ 2,\ 3$, and the random variables $X_\jmath, \jmath = 1, ..., 4$, obtained from (3.8). These results reveal that the head and forearm variables are strongly correlated with the first principal component. For the second principal component, the tail (at 0.643) has the highest correlation value, and the height variable is moderately negatively correlated (with a value of -0.474), while the head and tail variables are weakly correlated. Combining this result with the visualizations of Figure 14 (and Figure 15), we observe, e.g., that the species NOCT, MGES, and GMUR of group G_1 which are identified primarily by their first principal component are equivalently identified by variations in their head and forearm measurements, but not on their tail or height; while in contrast the group G_5 species BARB and OGRIS are distinguished more by their second principal component, i.e., by their tail and moderately by their height measurements with the contribution of their head and forearm low.

Table 16 - Vertices Method, Correlations $C_{j\nu}$ between X_j and PC_ν: Bats

X_j	$PC1$	$PC2$	$PC3$
Head	0.8699	0.1067	-0.4234
Tail	0.7097	0.6433	0.2869
Height	0.7926	-0.4739	0.3415
Forearm	0.9053	-0.1919	-0.1170

9.3 Classical Midpoint Surrogates

When the interval midpoints are used as a classical surrogate value for the symbolic interval values, a plot of the resulting first and second principal components produces that given in Figure 17. There are five or six distinct groups, viz., G'_1 = {SCOM, NOCT, MGES, GMUR}, G'_2 = {MNAT, MDEC, SBOR, SBIC, OCOM}, G'_3 = {MBEC, PRH, MDAUB, MGP}, G'_4 = {PIPC, MOUS, PIPS, PIPN}, G'_5 = {BARB, OGRIS, MSCH}, G'_6

= {FCHEV}. Comparable results hold when the interval endpoints are used as classical surrogates. Possibly the groups G'_2 and G'_4 can be re-combined to give a single core group, except for the species PIPC which forms a group on its own.

9.4 Comparison of Symbolic and Classical Analyses

An immediate observation is that the groupings identified by the two methods are different in some aspects. An obvious difference relates to the species SCOM. The classical analysis has this species well contained in group G'_1. On the other hand, the vertices symbolic analysis identifies this species as quite distinct from the other species of G'_1 and instead identifies it as part of the symbolic group G_2. The species MSCH is clearly identified as part of the symbolic grouping G_2, especially when $\alpha = 0.4$. This contrasts with the classical analysis which suggests this species belongs possibly to the group with the species BARB and OGRIS, but is otherwise ambivalent as to whether it belongs here in G'_5 or as its own unique group. Further, both the species SBIC and MGP are firmly embedded in the classical groupings (G'_2 and G'_3, repectively), whereas the symbolic analysis shows these species to be somewhat distinct from the other species in those groupings; see, especially Figure 14 when $\alpha = 0.5$.

These differences occur because the classical analysis is based on a single point, the interval midpoint, and not the entire interval as in the symbolic analysis. That is, the rest of the data values other than the midpoint (or the two endpoints for that classical surrogate) are ignored in the classical analysis; or, equivalently, the interval variations used in the symbolic analysis are ignored.

This utilization of the data information from the interval lengths also reveals itself further in the following observation. Figure 16, for the four bats BARB, OGRIS, GMUR, and MGES, also display the plot of the classical principal components (indicated by the asterisk $*$). As noted for the faces dataset, these classical values, as points in s (=2 here) dimensional principal component space cannot reflect the relative sizes of the p (=4 here) dimensional observation hypercube. Another observation is that the classical principal component point is not at the centroid of the symbolic principal component hypercube. This feature is because the observed values for different species have different interval lengths. The symbolic analysis utilizes all the variations collectively. If all observations had equal interval lengths, then the classical principal component point would coincide with the centroid of the symbolic principal component hypercube.

10 Conclusion

Symbolic data emerge in numerous ways in contemporary datasets. This work has focused on expanding the principal component methodology for classical data to the important new class of interval-valued data. The centers method is essentially an analysis between the observations while the vertices method is an analysis using both between and within observations variations. As illustrated herein, analyses of classical surrogates produce results that fail to capture all the variation inherent to the data.

Further, in contrast to previous contributions, the present work develops the concept of vertex contributions to the principal components, a concept not possible in a classical analysis. In addition, the present paper permits hypercubes to be of $q < p$ dimension, as can happen when a given variable assumes a classical rather than an interval value. Furthermore, we introduce general weight functions not considered elsewhere.

An expanded discussion of the present material can be found in Billard et al. (2007). That technical report also includes a third dataset which deals with a panel of $m = 4$ judges who rate the quality of $p = 6$ wines. A judge's rating represents a measure of uncertainty with shorter (longer) interval ratings reflecting a higher (lower) level of surety of a wine's quality. Also, through this example, it is shown that, while dimensionality problems persist for classical and centers principal component analyses when $m < p$, for the vertices method these dimensionality problems only occur when the number of vertices $n < p$. If there are no trivial intervals in the dataset, this becomes $n = m2^p < p$. A pedagogical treatment of some of the basics in Billard and Diday (2006) contains additional examples.

Given the inevitable continued growth in the size of datasets, it is important to develop methodologies for other classes of symbolic data such as multi-valued and histogram-valued data. There is also a need to develop theoretical underpinings to all these methods. These remain as outstanding problems for future researchers.

Finally, algorithms for executing the symbolic analyses for interval data are available on the SODAS1.4 webpage (http://www.ceremade.dauphine.fr/touati/sodaspagegarde.htm).

Acknowledgement:

Partial funding support from the National Science Foundation is gratefully acknowledged. The authors also thank the referees and Associate Editor for their very helpful suggestions which have improved the presentation.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.), John Wiley, New York.

Baron, R. J., (1981). Mechanisms of human facial recognition. *International Journal of Man Machine Studies* 15, 137-178.

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In: *Analysis of Symbolic Data* (eds. H.-H. Bock and E. Diday), Springer, 103-124.

Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association* 98, 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley, Chichester.

Billard, L., Douzal-Chouakria, A., and Diday, E. (2007). Principal Components for Intervals. On (http://www.stat.uga.edu/people/faculty/BILLARD/Lynne.html).

Cazes, P. Chouakria, A. Diday, E. and Schecktman, Y. (1997). Extension de l'analyse en composantes principales des donnes de type intervalle. *Revue Statistique Applique* 45, 5-24.

Chellappa, R., Wilson, C.L. and Sirohey, S. (1995). Human and machine recognition of faces, a survey. *Proceedings IEEE* 83, 705-740.

Chouakria, A. (1998). *Extension des mthodes d'analyse factorielle à des données de type intervalle*, These de doctorat., Universite Paris Dauphine, Paris.

Chouakria, A., Diday, E. and Cazes, P. (1995). Extension of the Principal Components Analysis to Interval Data. In: *New Techniques and Technologies for Statistics* (eds. W. Klosgen, P. Nanopoulos and A. Unwin), GMD Forschungszentrum Infornationstechnik, Germany.

Chouakria, A., Diday, E. and Cazes, P. (1998). An Improved Factorial Representation of Symbolic Objects. In: *Advances in Data Science and Classification* (eds. A. Rizzi, M. Vichi and H.-H. Bock), Springer-Verlag, Rome, 397-402.

Chouakria, A., Cazes, P. and Diday, E. (2000). Symbolic Principal Component Analysis, in *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H. -H. Bock and E. Diday), Springer-Verlag, Berlin, p 200-212.

Coppi, R., Giordani, P and D'Urso, P. (2006). Component Models for Fuzzy Data. *Psychometrika* 71, 733-761.

Craw, I., Tock, D. and Bennett, A. (1992). Finding face features. In *Proceedings of the Second European Conference on Computer Vision*, 92-96.

Craw, I. and Cameron, P. (1996). Face recognition by computer. *Proceedings British Machine Vision Conference*, 489-507.

De Carvalho, F. A. T. 1998. Extension based proximity coefficients between Boolean symbolic objects. In *Data Science, Classification, and Related Methods* (eds. C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y.Tanaka and Y. Baba), Springer-Verlag, Berlin, 370-378.

Deif, A.S. (1991). Singular Values of an Interval Matrix. *Linear Algebra and its Applications* 151, 125-133.

Denoeux, T. and Masson, M. H. (2004). Principal Component Analysis of Fuzzy Data Using Autoassociative Neural Networks. *IEEE Transactions on Fuzzy Systems* 12, 336-349.

D'Urso, P. and Giordani, P. (2004). A Least Squares Approach to Principal Component Analysis for Interval Valued Data. *Chemometrics and Intelligent Laboratory Systems* 70, 179-192.

Etemad, K. and Chellappa, R. (1997). Discriminant analysis for recognition of human fac images. *Journal of the Optical Society of America*, 14, 1724-1733.

Fischler, M.A. and Eschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, c-22, 67-92.

Gioia, F. and Lauro, C. N. (2006). Principal Component Analysis on Interval Data. *Computational Statistics* 21, 343-363.

Giordani, P. and Kiers, H. A. L. (2004). Principal Component Analysis of Symmetric Fuzzy Data. *Computational Statistics and Data Analysis* 45, 519-548.

Giordani, P. and Kiers, H. A. L. (2006). A Comparison of Three Methods for Principal Component Analysis for Fuzzy Interval Data. *Computational Statistics and Data Analysis* 51, 379-397.

Ichino, M. (1988). General Metrics for Mixed Features - The Cartesian Space Theory for Pattern Recognition. In: *Proceedings IEEE International Conference on Systems, Man and Cybernetics*, 1, 494-497.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5th ed.), Prentice Hall, New Jersey.

Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer-Verlag, New York.

Kass, M., Witkin, A. and Terzopoulos, D. (1987). Snakes: Active contour models. In *IEEE Proceedings of the International Conference on Computer Vision*, 259-268.

Lauro, C. N. and Palumbo, F. (2000). Principal Component Analysis of Interval Data: A Symbolic Analysis Approach. *Computational Statistics* 15, 73-87.

Lauro, C. N. and Palumbo, F. (2005). Principal Component Analysis for Non-Precise Data. In: *New Developments in Classification and Data* (eds. M. Vichi, P. Morani, S. Mignami and A. Montanari), Springer-Verlag, Berlin, 173-183.

Lauro, C. N. and Gioia, F. (2006). Dependence and Interdependence Analysis for Interval-Valued Variables. In: *Data Science and Classification*(eds. V. Batagelj, H.-H Bock, A. Ferligoj and A. iberna), Springer-Verlag, Berlin, 171-183.

Leroy, B., Chouakria, A., Herlin, I. and Diday, E. (1996). Approche géométrique et classification pour la reconnaissance de visage. *Reconnaissance des Forms et Intelligence Artificelle*, INRIA and IRISA and CNRS, France, p 548-557.

Li, B. and Chellappa, R. (2002). A generic approach to simultaneous tracking and verification in video. *IEEE Transactions Image Processes* 11, 530-544.

Li, S.Z. and Lu, J. (1999). Face recognition using the nearest feature line method. *IEEE Transactions Neural Networks* 10, 439-443.

Moghaddam, B. and Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Transactions Pattern Analysis Machine Intelligence* 19, 696-710.

Moore, R. E. (1966). *Interval Analysis*, Prentice Hall, New Jersey.

Moon, H. and Phillips, P.J. (2001). Computational and performance aspects of PCA-based face recognition algorithms. *Perception* 30, 301-321.

Palumbo, F. and Lauro, C. N. (2003). A PCA for Interval Valued Data Based on Midpoints and Radii. In: *New Developments in Psychometrics* (eds. H. Yanai, A. Okada, K. Shigematu, Y. Kano and J. J. Meulman), Springer-Verlag, Japan, 641-648.

Rhon, J. (1993).Interval Matrices : Singularity and Real Eigenvalues. *SIAM Journal Matrix Analysis and Applications* 14, 82-91.

Samal, A. and Iyengar, P. (1992). Automatic recognition and analysis of human faces and facial expressions. A survey. *Pattern Recognition* 25, 65-77.

Staib, L.H. and Duncan, J.S. (1992). Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 1061-1075.

Turk, M. (1991). *Interactive-time vision: face recognition as a visual behavior.* PhD thesis, Massachusetts Institute of Technology.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3, 72-86.

Yabuuchi, Y., Watada, J., Nakamori, Y., 1997. Fuzzy Principal Component Analysis for Fuzzy Data. In: *Proceedings Sixth IEEE International Conference on Fuzzy Systems* 2, 1127 1132.

Zhao, W., Chellappa, R., Phillips, P.J. and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys* 35, 399-459.

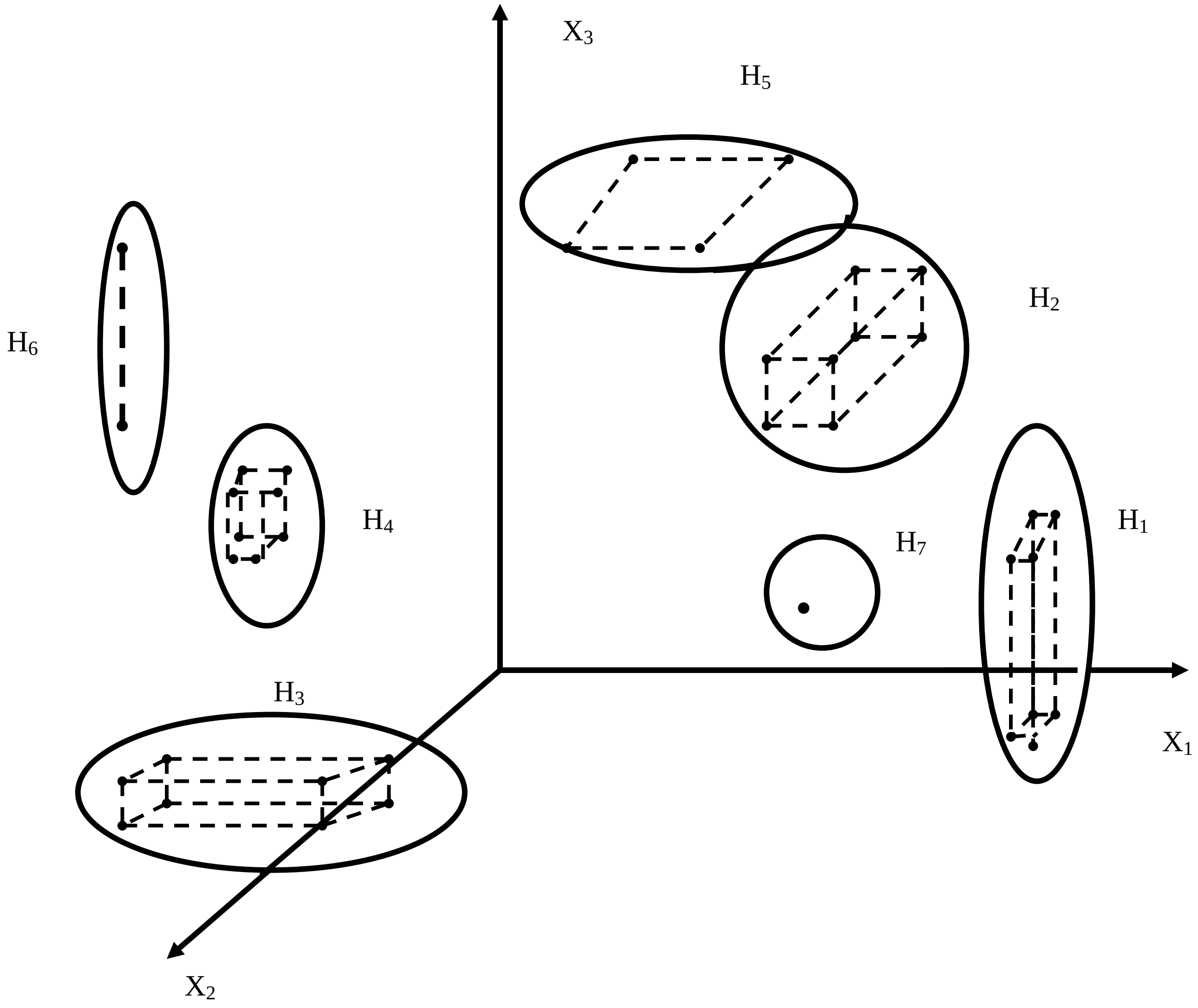

Figure 1 - Types of Hypercubes: Clouds of Vertices

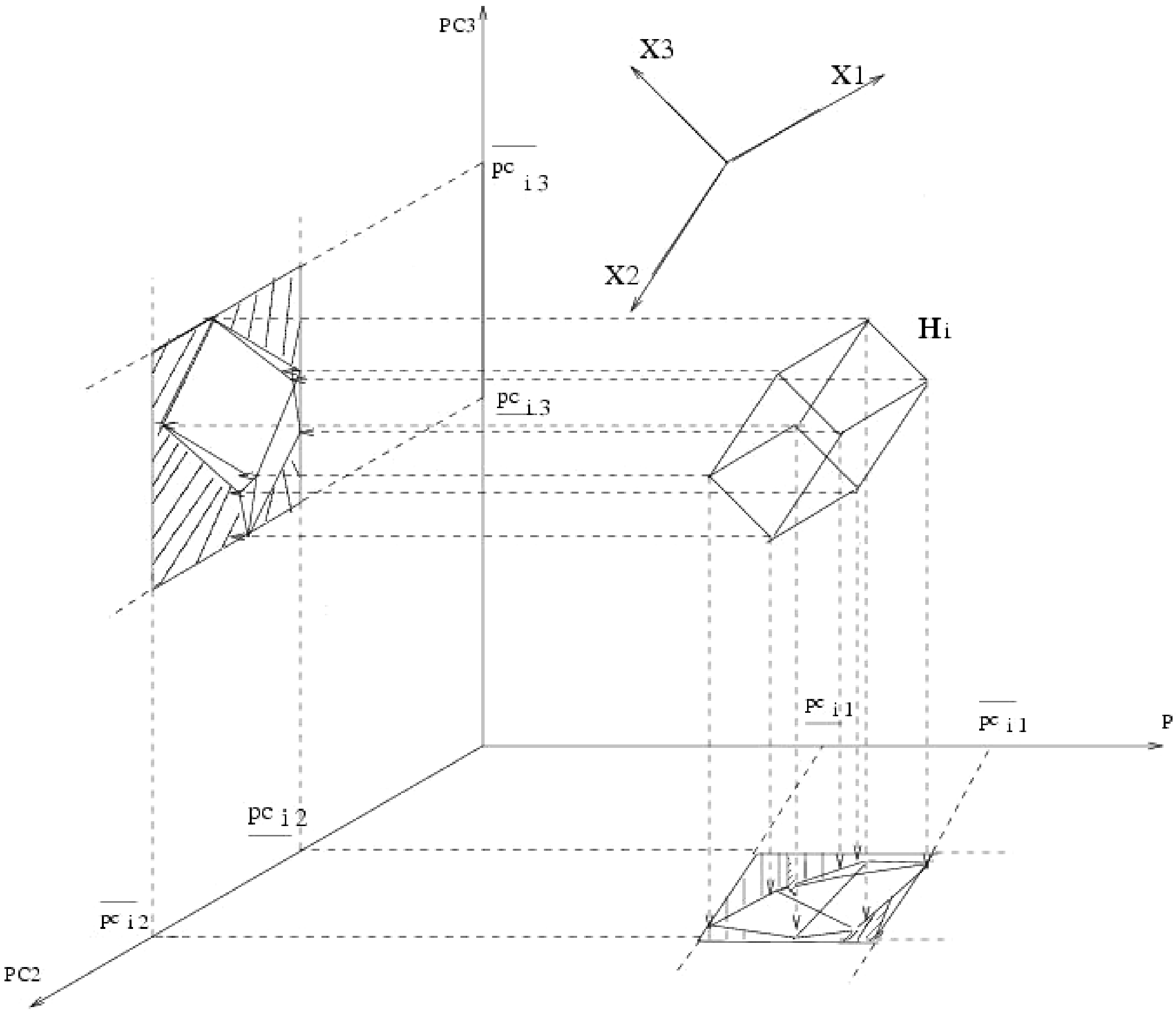
PC3
X3
X1
X2
Hi
$\overline{\text{pc}}_{i3}$
$\underline{\text{pc}}_{i3}$
$\underline{\text{pc}}_{i1}$
$\overline{\text{pc}}_{i1}$
PC1
$\underline{\text{pc}}_{i2}$
$\overline{\text{pc}}_{i2}$
PC2

Figure 2 - Projection Hypercube H_u to Principal Component ($\nu = 1, 2, 3$) Axes
($\overline{pci\nu} \equiv y^a_{i\nu}$, $\underline{pci\nu} \equiv y^b_{i\nu}$)

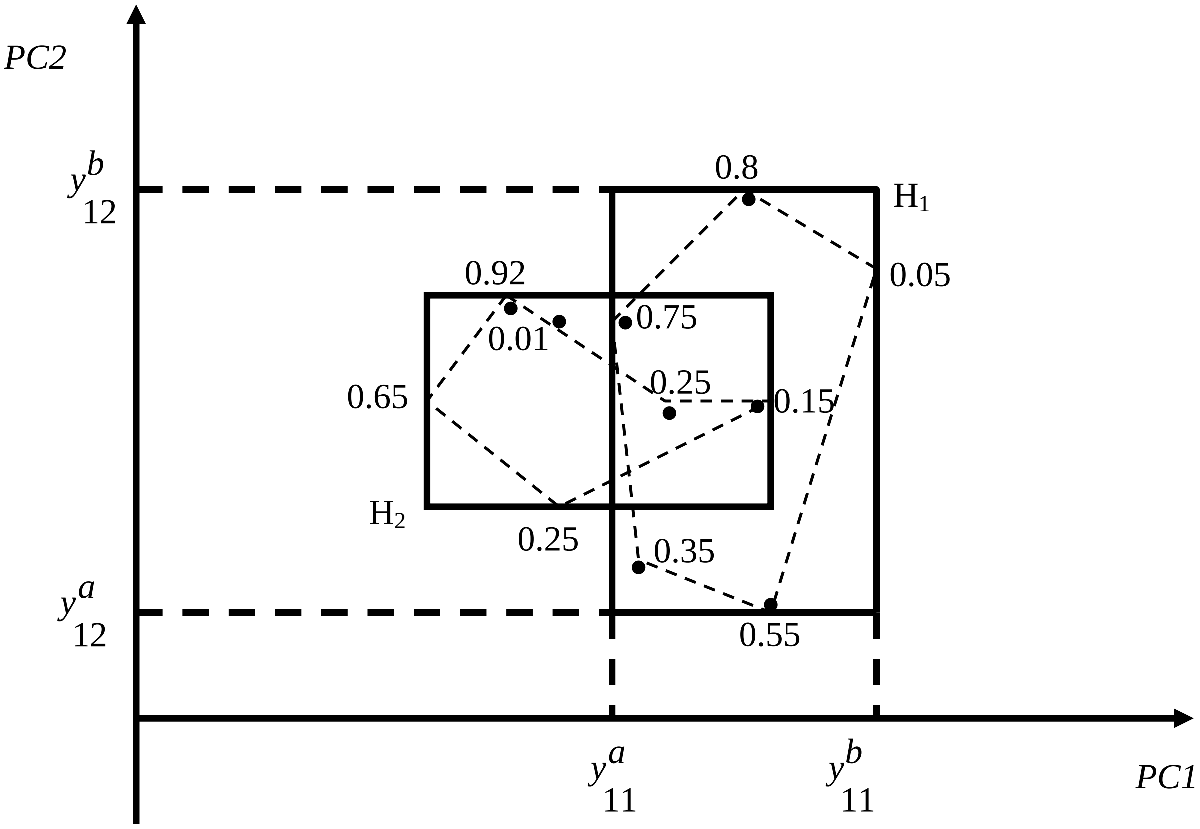

Figure 3 - Principal Component Envelope, $\alpha = 0$, based on Relative Contributions of Vertices to $PC\nu$, $\nu = 1, 2$

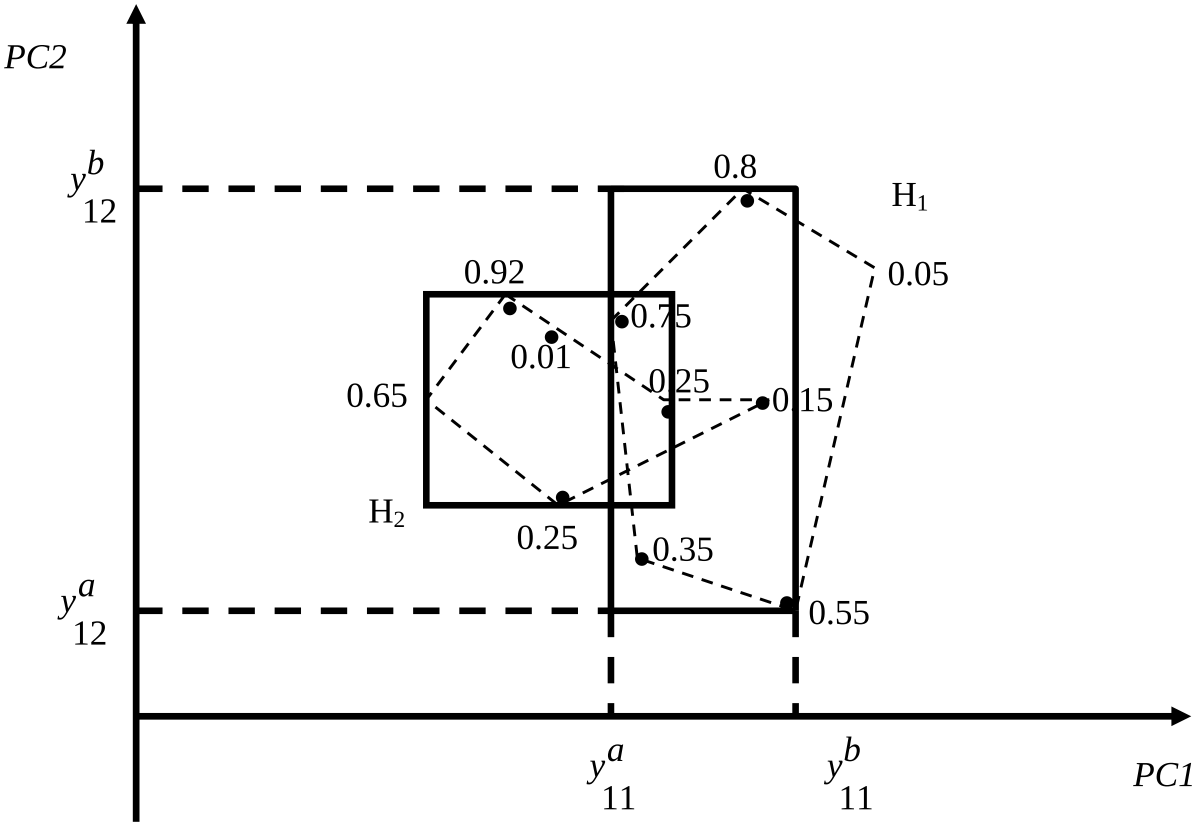

Figure 4 - Principal Component Envelope, $\alpha = 0.2$, based on Relative Contributions of Vertices to $PC\nu$, $\nu = 1, 2$

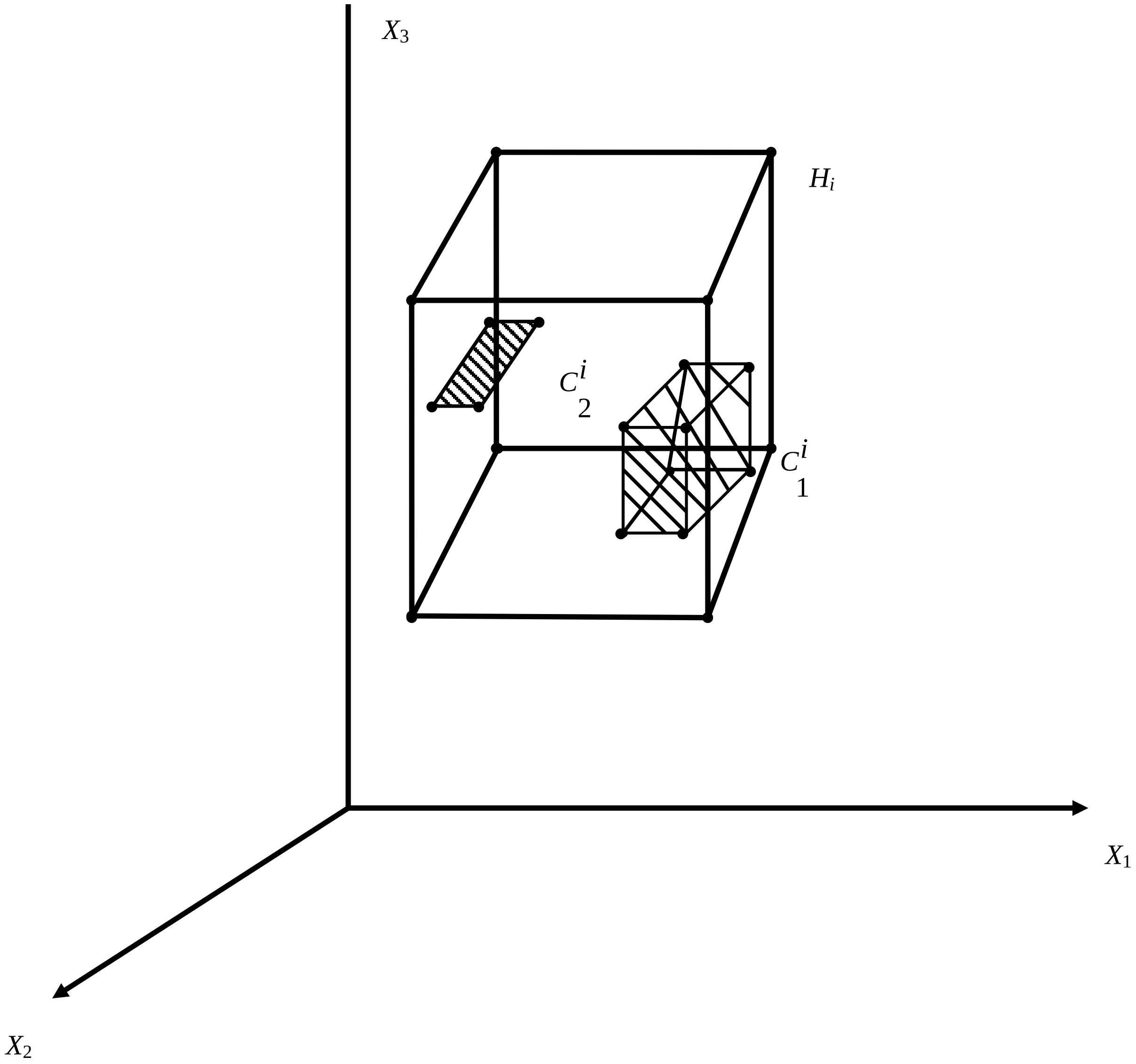

Figure 5 - Constrained Hypercubes: "Holes" C^i inside Data H^i

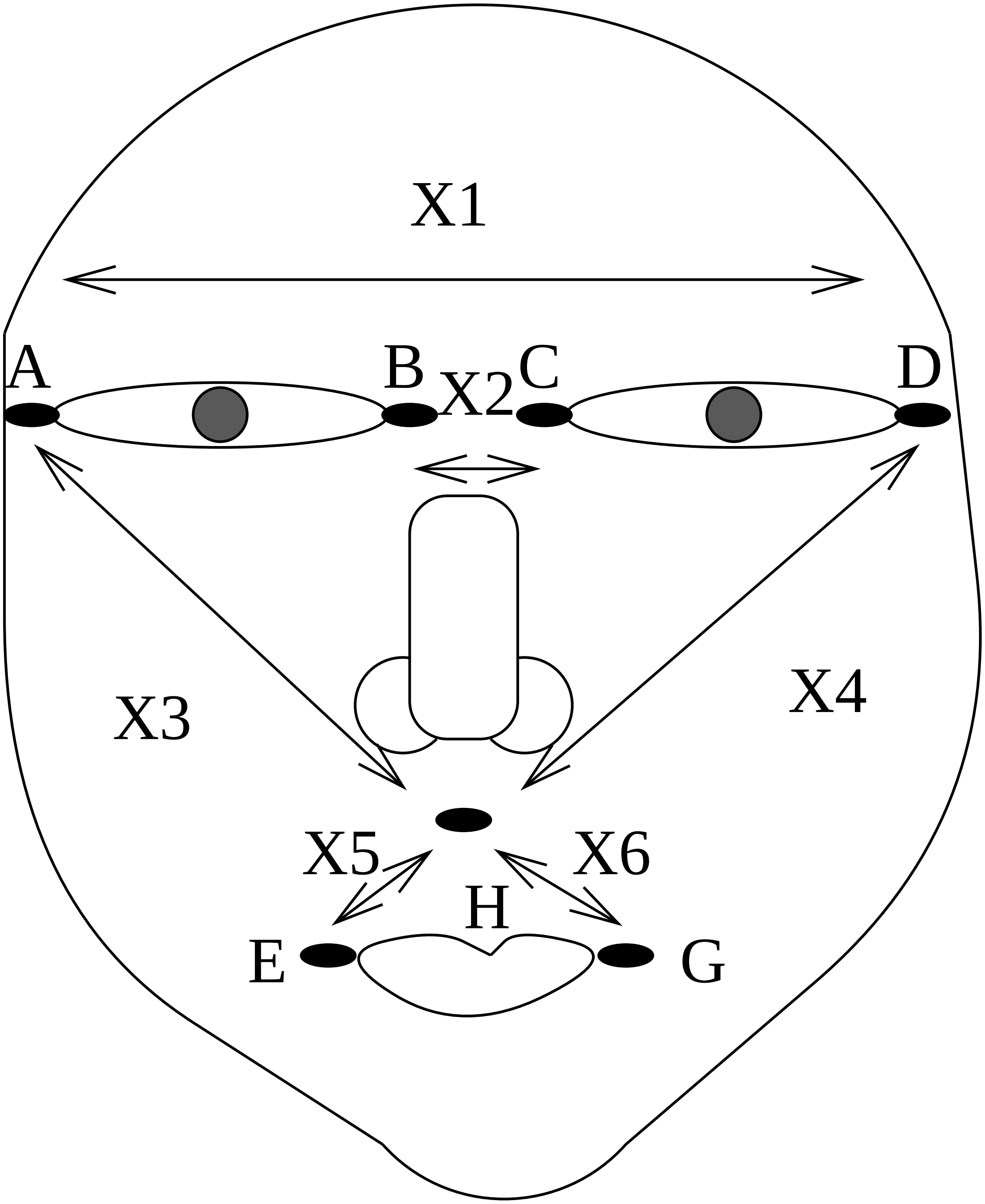

Figure 6 - Face: Description of Random Variables

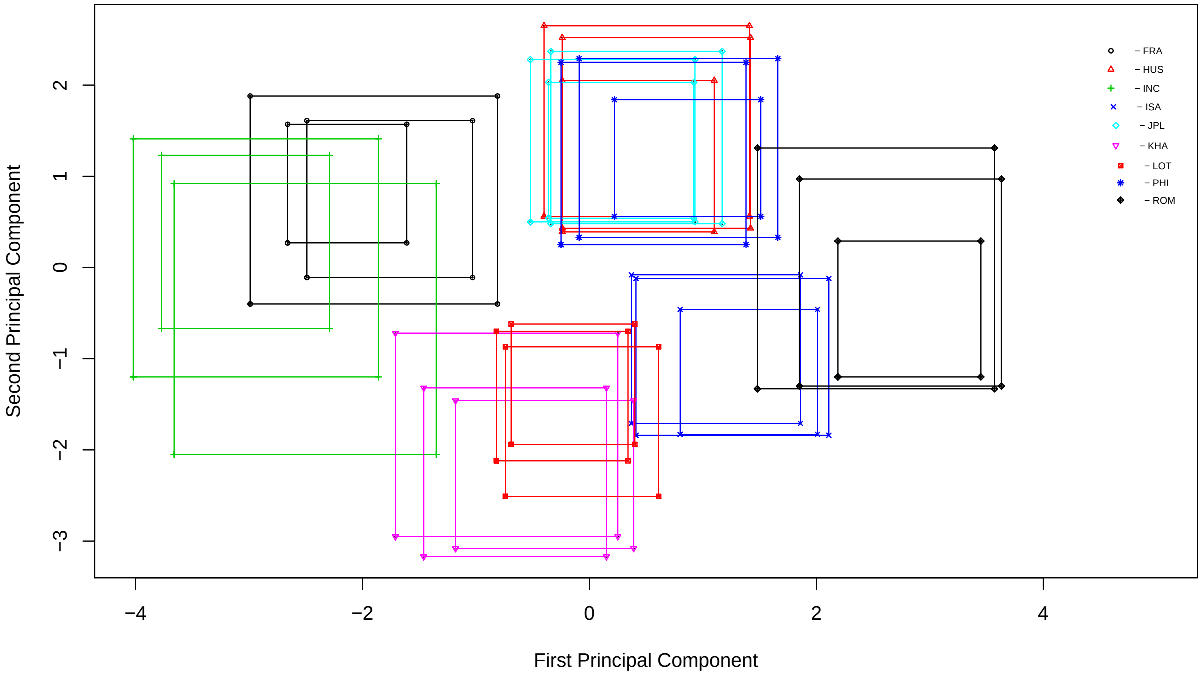

Figure 7 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$

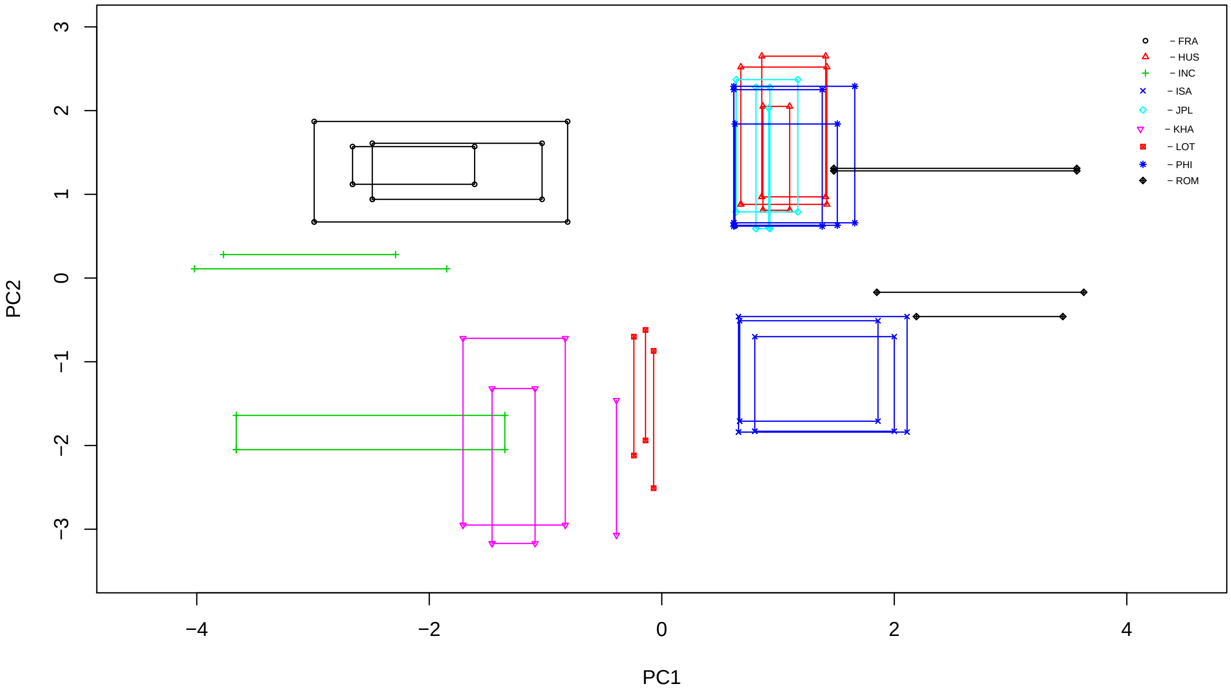

Figure 8 - Faces: Vertices Principal Components $PC\nu$, $\nu = 1, 2$: $\alpha = 0.2$

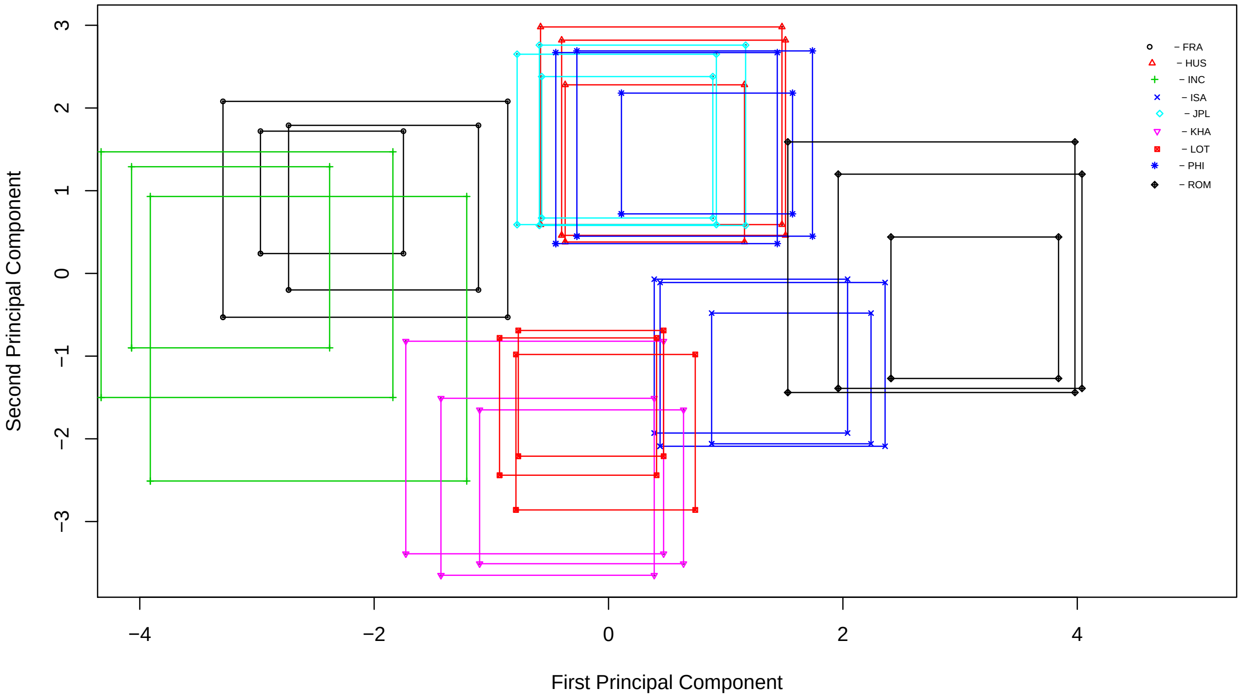

Figure 9 - Faces: Centers Principal Components $PC\nu$, $\nu = 1, 2$

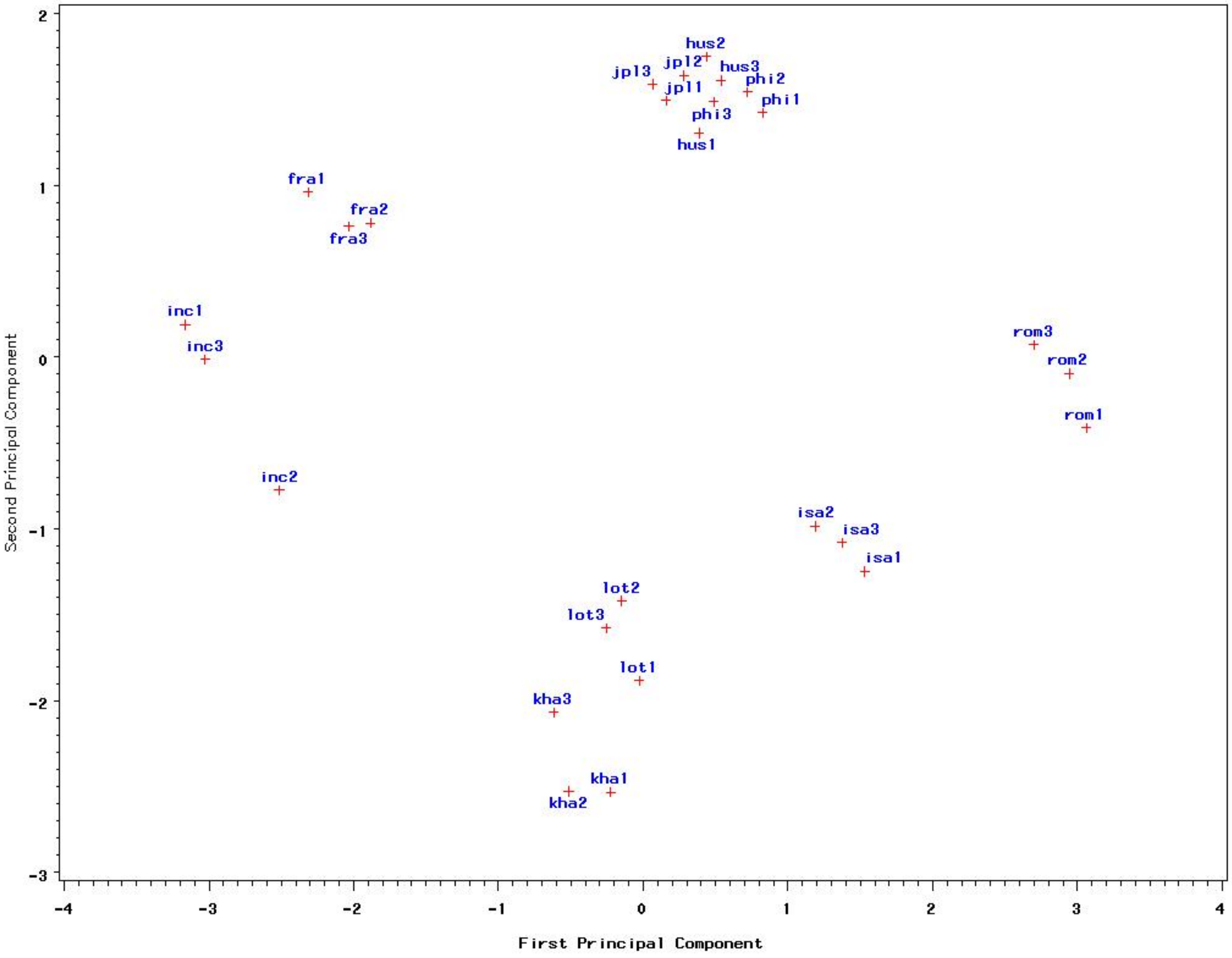

Figure 10 - Faces: Classical Principal Components $PC\nu$, $\nu = 1, 2$ - Midpoints

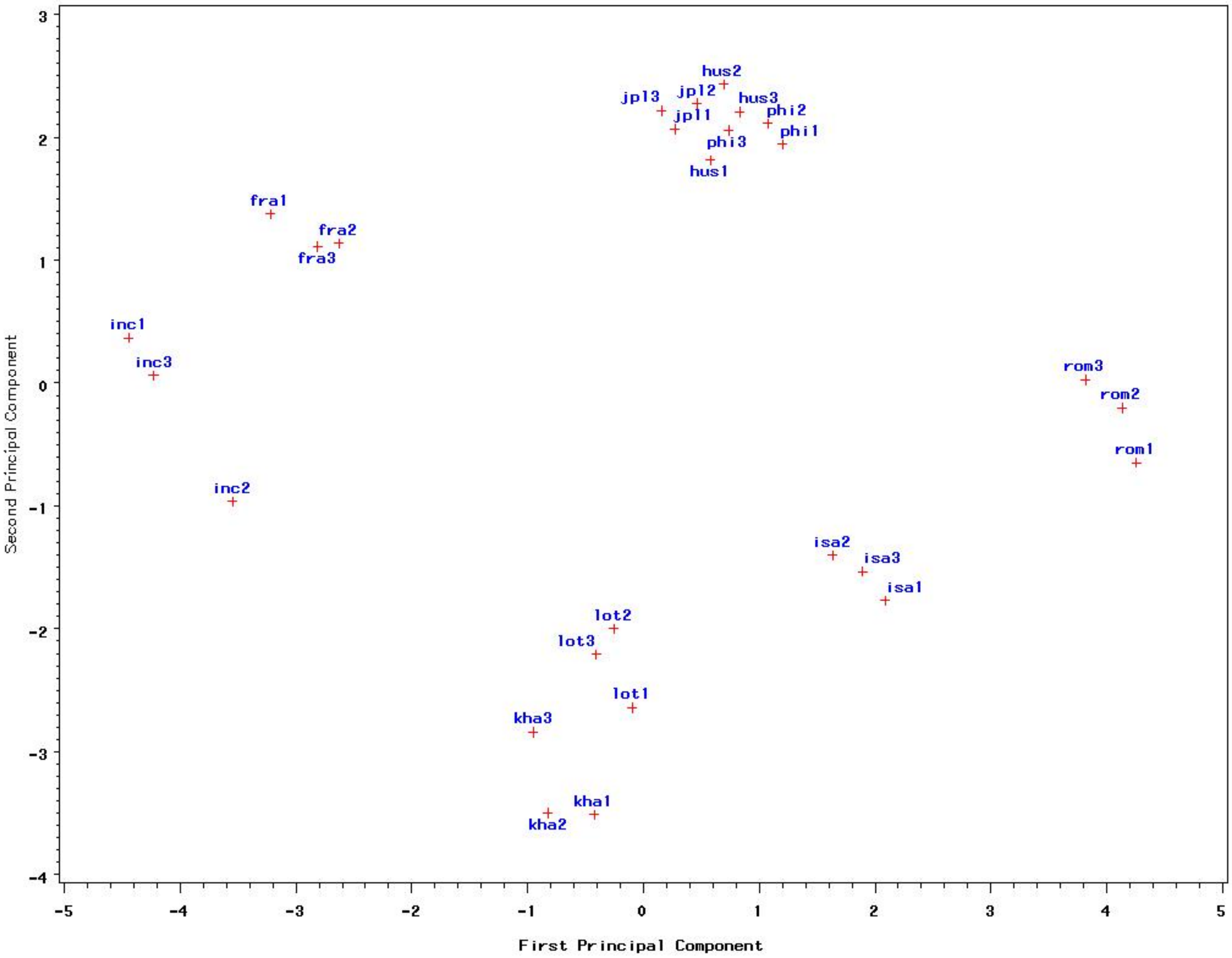

Figure 11 - Faces: Classical Principal Components $PC\nu$, $\nu = 1, 2$ - Endpoints

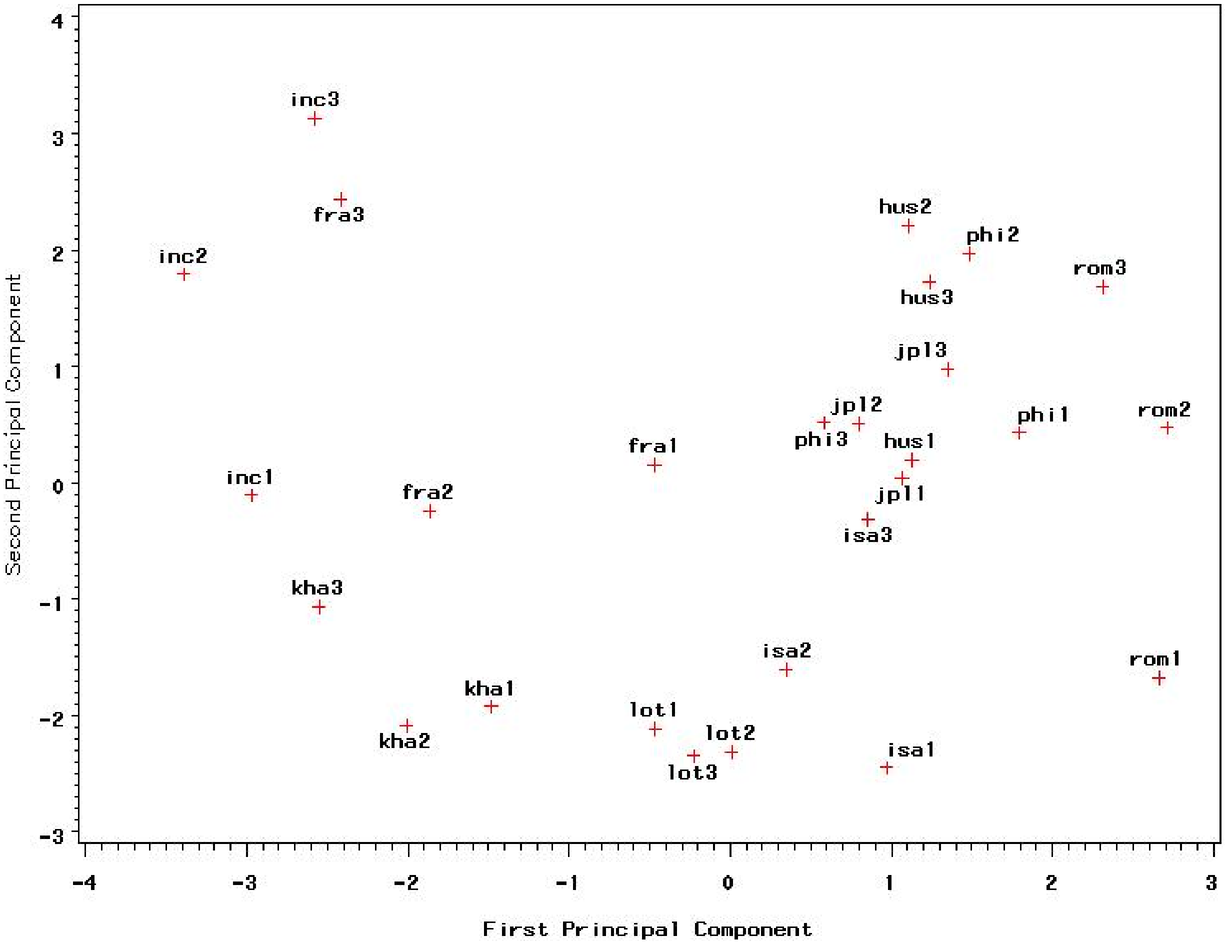

Figure 12 - Faces: Classical Principal Components $PC\nu$, $\nu = 1, 2$ - Ranges and Midpoints

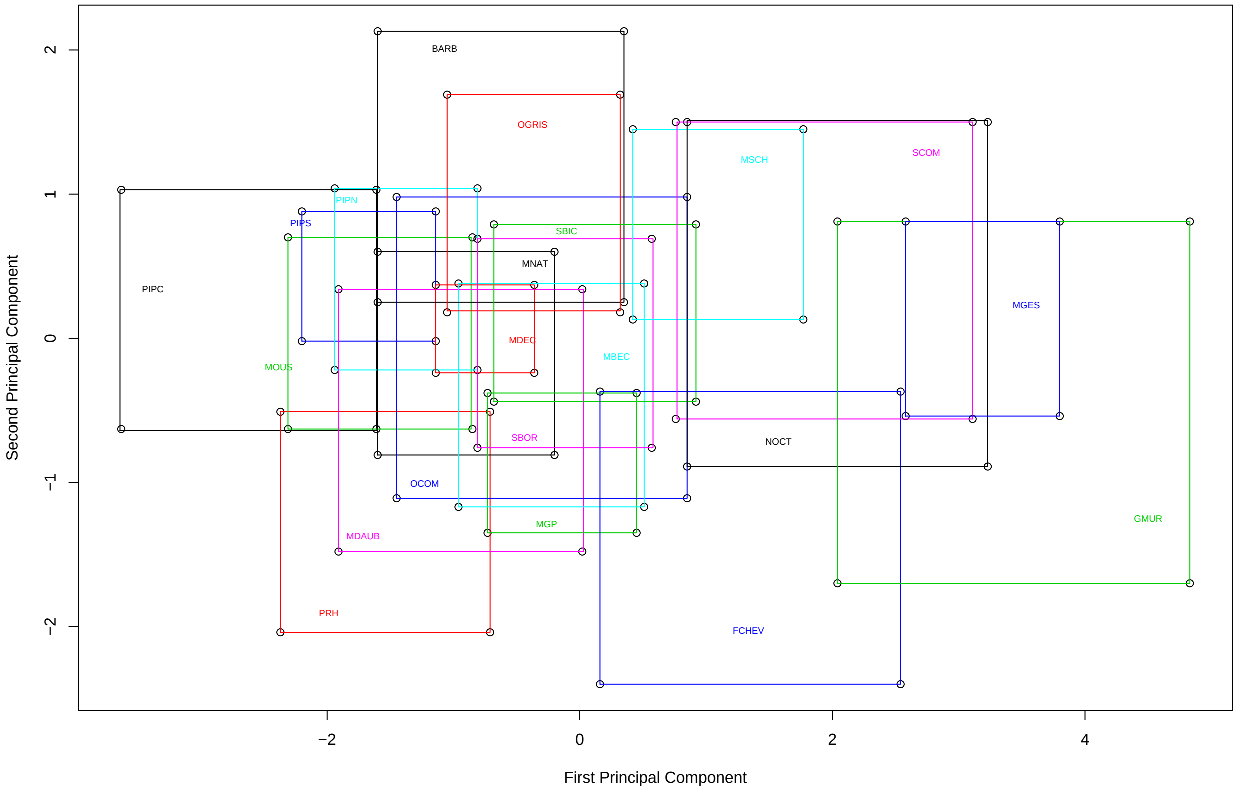

Figure 13 - Bats: Vertices Principal Components $PC\nu$, $\nu = 1, 2$

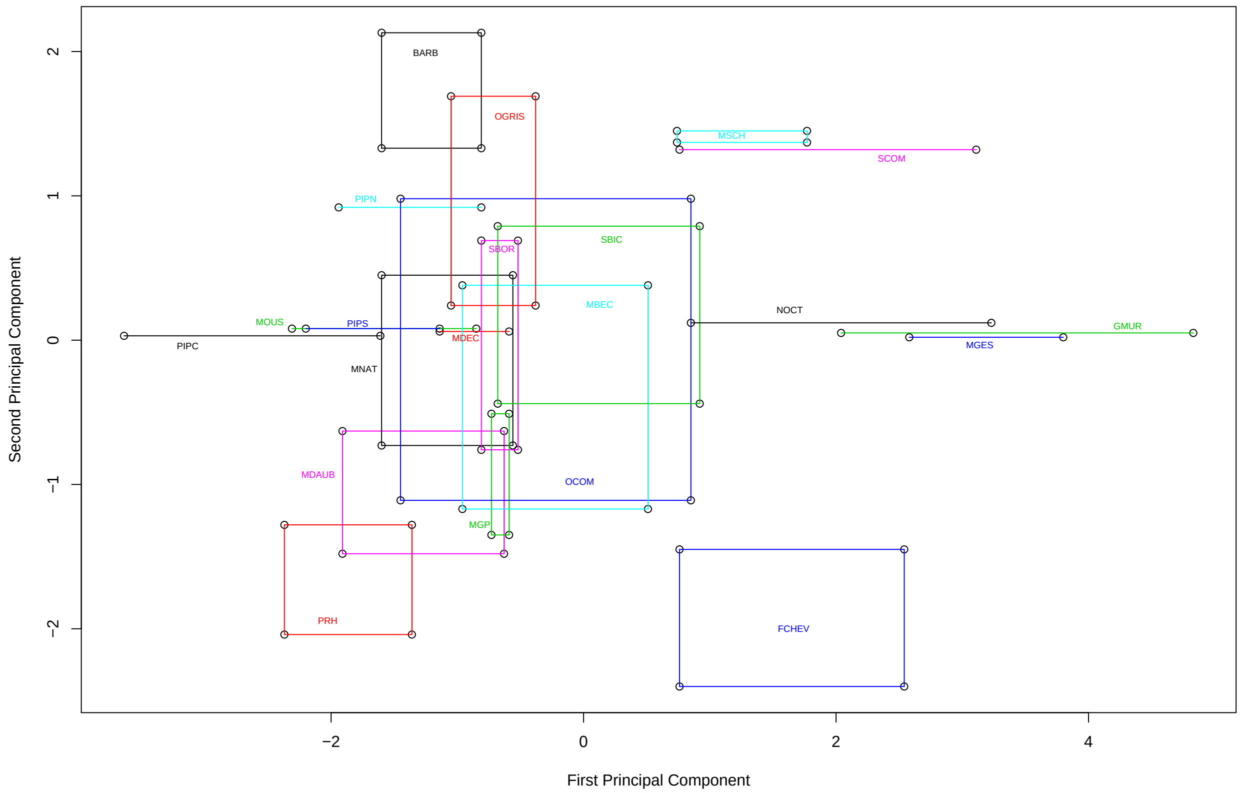

Figure 14 - Bats: Vertices Principal Components $PC\nu$, $\nu = 1, 2$ - $\alpha = 0.4$

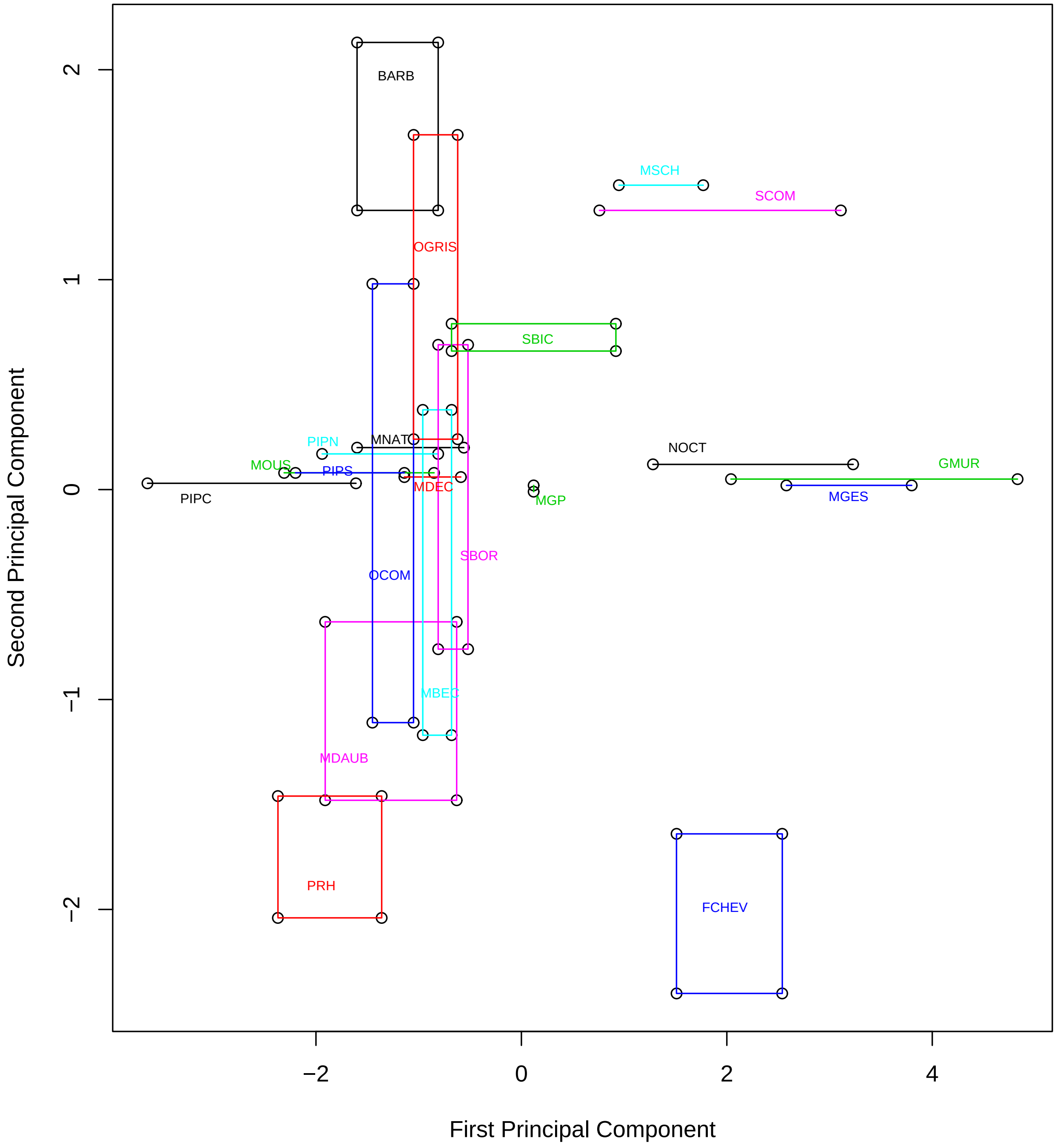

Figure 15 - Bats: Vertices Principal Components $PC\nu$, $\nu = 1, 2$ - $\alpha = 0.5$

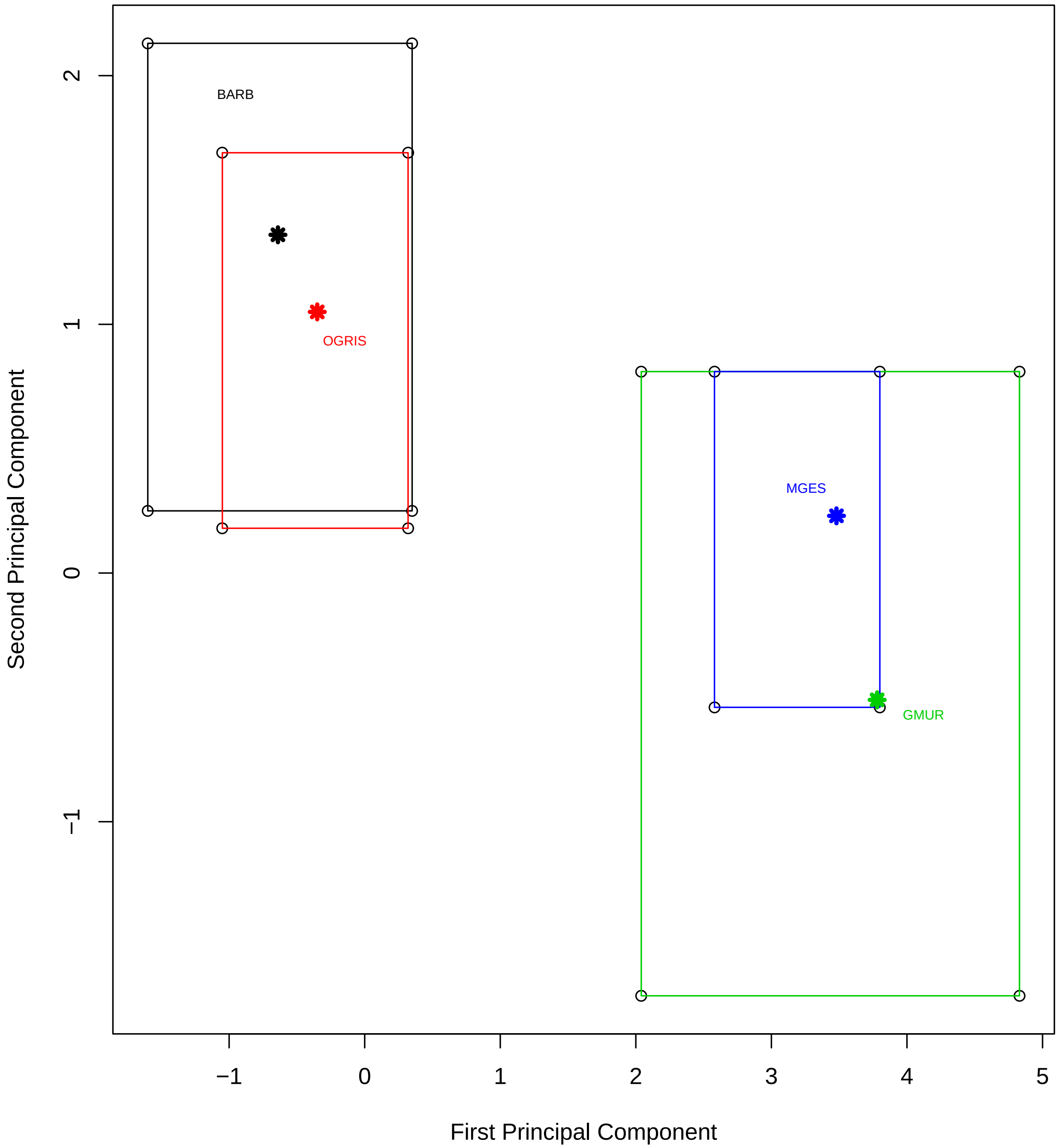

Figure 16 - Bats BARB, OGRIS, GMUR, MGES: Principal Components $PC\nu$, $\nu = 1, 2$

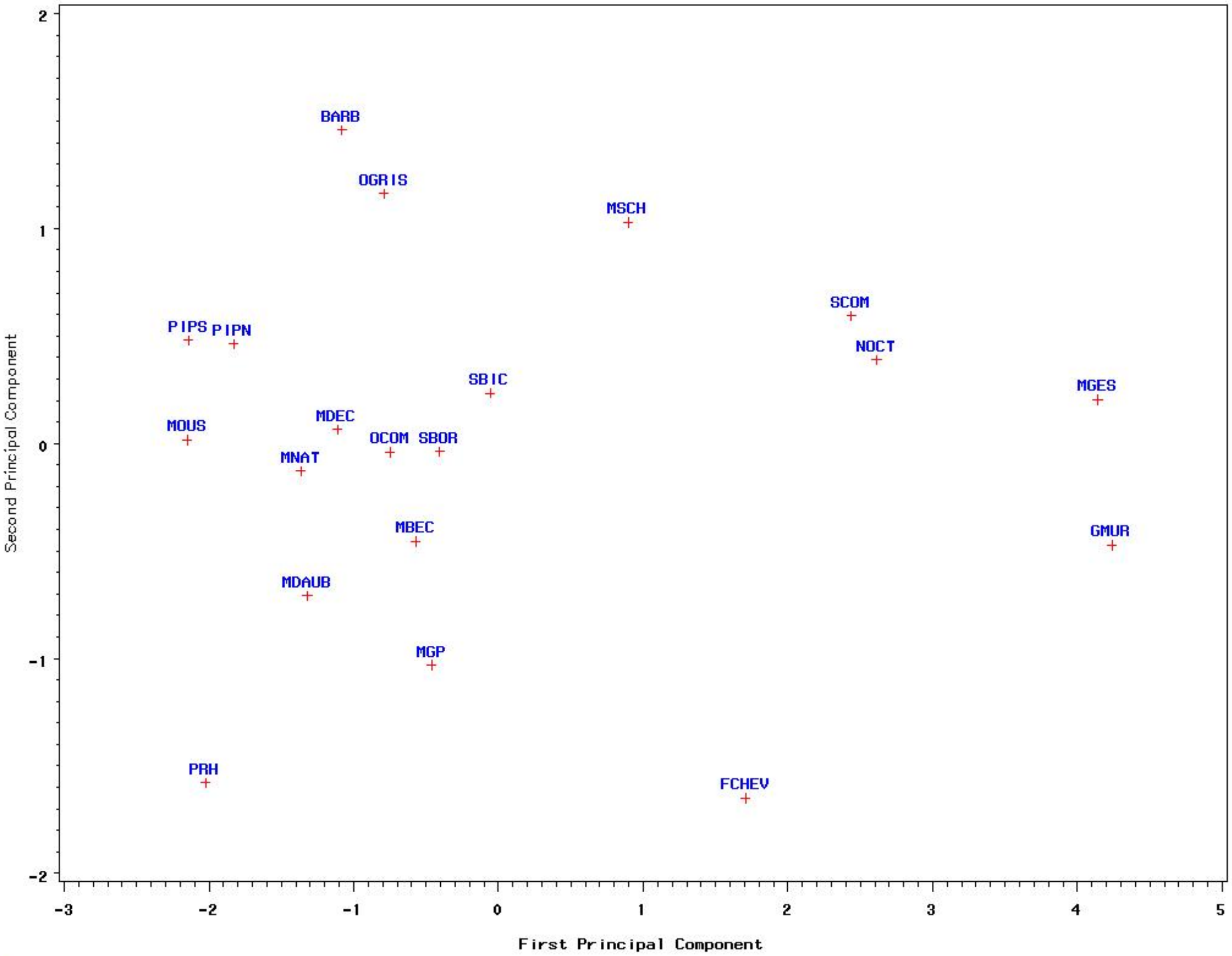

Figure 17 - Bats: Classical Principal Component $PC\nu$, $\nu = 1, 2$ - Midpoints