Histogram Algebra: an Application to Histogram Principal Component Analysis

Eduardo Mendes
eduardo.mendes@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2019

Abstract

Symbolic Data Analysis has become an increasingly important area of Statistics over the past few years, due to the increasing data complexity to take into account. One of the most used types of symbolic variables in this area is the histogram, which contains information about the probability distribution of the individuals that originated it. This is the reason why it is important to find ways to easily manipulate and do arithmetic operations with this type of variable. As interval variables are a particular type of histogram variables, they also deserve special attention in this work. After focusing on the creation of an algebra for histograms, based on the arithmetic operations with quantile functions, which is one of the possible ways to represent histograms with, a general expression for the computation of linear combinations between histograms is obtained. The above mentioned expression is then used to propose a new estimation method for Symbolic Principal Components Analysis, when it is applied to histogram-valued data. The output of the method is a multivariate histogram representation of the original observation in a space spanned by the Symbolic Principal Components, unlike what happens in the vast majority of the works in this area. This method is applied to two data sets. Based on the results, its advantages and limitations are also analyzed.
Keywords: Symbolic Data Analysis, histogram-valued variables, quantile functions, histogram algebra, symbolic covariance, Symbolic Principal Components

1. Introduction

The ever-growing importance of Symbolic Data Analysis, which allows us to analyze data with inherent variability, created the need to find better and easier ways to manipulate this kind of data. One of the types of symbolic variables that has gained increasing importance over the past years is the histogram-valued variable. It represents data as histograms, which contain information about the probability distribution of the individuals that originated them. The arithmetic with quantile functions introduced in [5] is one of the best known methods to perform mathematical operations with histograms. However, these operations have not yet been generalized as an algebra. The first goal of this work is to define this algebra and find a general expression which would make it possible to compute linear combinations of histograms easily. This could be useful for many statistical methods that use histogram variables, as it is the case of the Symbolic Principal Components Analysis (SPCA). In the vast majority of the works in this area, the Principal Component scores, which are the end result of this method, are not represented as histograms. As this is not ideal in many cases, our second goal is to define histogram-valued Principal Components with the help of the generalized expression for the linear combination of histograms deduced in this work. This helps us to improve the SPCA estimation methods applied to histogram-valued variables, allowing SPCA to be used as a dimensionality reduction technique, which is useful when combined with other statistical methods.

2. Interval and Histogram data

The main focus of this work is related to interval and histogram-valued variables. Firstly, it is important to define what a symbolic variable is (see [2], Chapter 3).

Definition 2.1. *A symbolic variable X_l is a mapping from a set E of statistical entities, such that:*

$$X_l : E \to B$$
$$X_l(e_j) = \epsilon_{jl}, \forall e_j \in E.$$

The statistical entities e_j from E (the individuals from a population) can be originated either from single individuals (micro-data) or from a collection of micro-data whose information was aggregated

to create more complex symbolic objects (macro-data). Each observation j of a variable takes its value from the set B, which varies according to the type of symbolic variable we are dealing with. The result $\{X_{jl} = \epsilon_{jl}\}$ represents the symbolic value that the variable l takes for the observation j. Taking this into consideration, we can define histogram-valued variables in the following way:

Definition 2.2. *A histogram-valued variable, X, corresponds to a transformation from the set of entities that defines the population, E, into a set B of possible histograms. A histogram x_j is a set of subintervals, $\mathcal{Y}_j = (y_{1j}, .., y_{n_j j})^t$ and a set of probabilities associated to each subinterval, $\mathcal{P}_j = (p_{1j}, .., p_{n_j j})^t$, where $\sum_{i=1}^{n_j} p_{ij} = 1$, p_{ij} is the probability of the j-th entity assuming a value in $y_{ij} = [a_{ij}, b_{ij}[$, where $a_{ij} \leq b_{ij}$, $a_{i+1j} = b_{ij}$, and n_j is the total number of subintervals associated with the j-th entity.*

Interval-valued variables can be seen as a particular case of a histogram-valued variable where $n_j = 1$.

It is also relevant to mention some different ways that are used to represent intervals and histograms. For instance, intervals are generally represented using a notation with their lower and upper bounds:

$$y = [a, b] \quad \text{with } a, b \in \mathbb{R};\ a \leq b, \tag{1}$$

where a is the lower bound and b the upper bound of the interval y.

Histograms are characterized by a set of subintervals, such that the i-th subinterval of x_j is represented as [a_{ij}, b_{ij}[, with the associated probability p_{ij}. Admitting this notation, a histogram x_j with n_j subintervals can be represented as

$$\begin{aligned} x_j = \{&[a_{1j}, b_{1j}[, p_{1j}; [a_{2j}, b_{2j}[, p_{2j}; ...; \\ &; [a_{ij}, b_{ij}[, p_{ij}; ...; [a_{n_j j}, b_{n_j j}], p_{n_j j}\}, \end{aligned} \tag{2}$$

with $a_{ij}, b_{ij} \in \mathbb{R}$, $a_{ij} \leq b_{ij}$, $a_{i+1j} = b_{ij}$, $0 \leq p_{ij} \leq 1$, and $\sum_{i=1}^{n_j} p_{ij} = 1$; $i \in \{1, .., n_j\}$.

An alternative method of representation for intervals and histograms is to use the center and range of each interval/subinterval. The center, c_y, of an interval $y = [a, b]$ is given by the expression $c_y = \frac{a+b}{2}$ and the respective range, r_y, is given by $r_y = b - a$.

Since $b \geq a$ is always a condition for an interval, the range is always a non-negative value, while the center can take any value in $\mathbb{R}$.

Using the previous definitions of center and range of an interval, it is possible to characterize an interval only as a set of its center and range:

$$y = (c_y, r_y), \quad \text{with } c_y \in \mathbb{R}, r_y \in \mathbb{R}_0^+. \tag{3}$$

Similarly to (3), a histogram can be represented as a set of three vectors related to the center, ranges, and associated probabilities of its subintervals ($\boldsymbol{c}_j$, $\boldsymbol{r}_j$, and $\boldsymbol{p}_j$, respectively). Therefore, for a histogram x_j, the representation of its centers, ranges, and probabilities is:

$$x_j = (\boldsymbol{c}_j, \boldsymbol{r}_j, \boldsymbol{p}_j). \tag{4}$$

Another useful way to represent intervals and histograms is through a quantile function, which is related to the cumulative distribution function (cdf), F_X, of a random variable X. A cdf F_X is given by the expression $F_X(x) = P(X \leq x)$, with $x \in \mathbb{R}$.

The quantile function can be seen as the inverse of F_X. This concept was introduced in [5] for histogram variables. It gives the value x for which the probability of the random variable being less than or equal to that same x is equal to a given value p ($0 \leq p \leq 1$). This is equivalent to finding the minimum value of x among all the values whose cdf value exceeds p. This value p must always lie between 0 and 1, since it corresponds to a probability. Thus, a quantile Q_X function for a random variable X with cdf $F_X(x)$ can be defined by

$$Q_X(p) = inf\{x \in \mathbb{R} : p \leq F_X(x)\},\ 0 \leq p \leq 1.$$

We assume that the micro-data associated with each interval and subinterval studied in this work follows a uniform distribution. Thus, it is possible to use the general expression of the cdf for continuous uniformly distributed random variables to characterize them. Under these conditions, the quantile function, Q_Y, of a random variable Y with a uniform distribution defined over an interval $[a, b]$ is

$$Q_Y(p) = a + (b - a)p,\ 0 \leq p \leq 1. \tag{5}$$

The expression (5) can also be applied to each of the subintervals of a histogram, since it is assumed that micro-data follows a uniform distribution within a subinterval. Hence, the quantile function, Q_X, of a histogram-valued variable X with n subintervals is:

$$Q_X(p) = \begin{cases} a_1 + \frac{p}{w_1}(b_1 - a_1), & 0 \leq p < w_1 \\ a_2 + \frac{p - w_1}{w_2 - w_1}(b_2 - a_2), & w_1 \leq p < w_2 \\ \vdots & \\ a_i + \frac{p - w_{i-1}}{w_i - w_{i-1}}(b_i - a_i), & w_{i-1} \leq p < w_i \\ \vdots & \\ a_n + \frac{p - w_{n-1}}{1 - w_{n-1}}(b_n - a_n), & w_{n-1} \leq p \leq 1 \end{cases}, \tag{6}$$

where the w_is are the cumulative probabilities of the first i subintervals and are defined by

$$w_i = \begin{cases} 0, & i = 0 \\ \sum_{k=1}^{i} p_k, & i = 1, .., n \end{cases}. \tag{7}$$

The quantile functions are always non-decreasing in their domain $[0, 1]$. This happens because the subintervals that are a part of a histogram are always consecutive and disjoint. That is, the lower bound of a subinterval i of a histogram is always greater than or equal to its respective upper bound ($b_i \geq a_i$), and the lower bound of the next subinterval is equal to the upper bound of the previous subinterval ($a_{i+1} = b_i$). The quantile function can also be non-continuous in some cases.

3. Interval Algebras

To better understand the histogram algebra that we are trying to deduce, it is first necessary to study the properties of some interval algebras. Two of the most relevant interval algebras are Moore's Interval Algebra and the Extended Interval Algebra, which are analyzed in this section.

3.1. Moore's Interval Algebra

The most commonly used algebra when dealing with interval variables is the one that was defined by Moore in [9]. This algebra will henceforth be denominated as Moore's Interval Algebra. The operations of this algebra, using the notation with the resulting centers and ranges of the intervals in (3), is summarized in Table 1.

Table 1: Summarized results for the operations of Moore's Interval Algebra.

Operation	Center	Range
$y_1 + \beta$	$c_1 + \beta$	r_1
$y_1 + y_2$	$c_1 + c_2$	$r_1 + r_2$
$y_1 - y_2$	$c_1 - c_2$	$r_1 + r_2$
βy_1	βc_1	$\lvert\beta\rvert r_1$
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\lvert\beta_1\rvert r_1 + \lvert\beta_2\rvert r_2$

In all the operations of Moore's Interval Algebra presented in Table 1 the range is always positive. This causes a disadvantage: the ranges of the resulting intervals keep increasing with each consecutive operation. As a consequence, very large intervals can be generated and, thus, lose their significance. It is also interesting to note that, for this reason, this algebra cannot be considered a vector space, as the additive inverse axiom fails to hold.

A generalized formula for the linear combinations with Moore's Interval Algebra of k intervals can be deduced. This is achieved by aggregating in vectors the sets of the centers and ranges of the corresponding intervals, and also the scalars β. In this way, the vectors $\boldsymbol{y} = (y_1, .., y_k)^t$, $\boldsymbol{c} = (c_1, .., c_k)^t$, $\boldsymbol{r} = (r_1, .., r_k)$ and $\boldsymbol{\beta} = (\beta_1, .., \beta_k)$ are obtained. Using these vectors, the general expression for the linear combination of intervals with Moore's Interval Algebra is given by:

$$\boldsymbol{\beta}^t \boldsymbol{y} = [\boldsymbol{\beta}^t \boldsymbol{c} - |\boldsymbol{\beta}|^t \frac{\boldsymbol{r}}{2},\ \boldsymbol{\beta}^t \boldsymbol{c} + |\boldsymbol{\beta}|^t \frac{\boldsymbol{r}}{2}], \tag{8}$$

where $\boldsymbol{\beta}^t$ is the transpose of $\boldsymbol{\beta}$ and $|\boldsymbol{\beta}|$ corresponds to the computation of the absolute values of all the elements in $\boldsymbol{\beta}$.

3.2. Extended Interval Algebra

There is another algebra for intervals that overcomes the issue of the increasing expansion of the ranges of the intervals with consecutive operations that occurs in Moore's Algebra. This algebra was defined in [6] and will henceforth be designated by Extended Interval Algebra.

The resulting centers and ranges of the intervals for all the operations in this algebra are presented in Table 2. From its analysis, it can be concluded that, while the resulting centers are the same as in Moore's, now the ranges can have a negative value, which is indicative of an extended interval where $b < a$. For instance, the extended interval $[-2, -4]$ has a range equal to -2. Hence, for this algebra, the difference between two equal intervals is now equal to 0, and all the axioms of a vector space are satisfied. Thus, the Extended Interval Algebra consists of a vector space.

Table 2: Summarized results for the operations of the Extended Interval Algebra.

Operation	Center	Range
$y_1 + \beta$	$c_1 + \beta$	r_1
$y_1 + y_2$	$c_1 + c_2$	$r_1 + r_2$
$y_1 - y_2$	$c_1 - c_2$	$r_1 - r_2$
βy_1	βc_1	βr_1
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\beta_1 r_1 + \beta_2 r_2$

It is also possible to obtain a general expression for the linear combination of intervals with the Extended Interval Algebra, by using the vectors $\boldsymbol{y}$, $\boldsymbol{c}$, $\boldsymbol{r}$, and $\boldsymbol{\beta}$, previously defined for (8):

$$\boldsymbol{\beta}^t \boldsymbol{y} = [\boldsymbol{\beta}^t \boldsymbol{c} - \boldsymbol{\beta}^t \frac{\boldsymbol{r}}{2},\ \boldsymbol{\beta}^t \boldsymbol{c} + \boldsymbol{\beta}^t \frac{\boldsymbol{r}}{2}]. \tag{9}$$

While this algebra was used with success in some works, it is hard to grasp, under this notation, what the interpretation of an interval $y = [a, b]$ with $a > b$ would be. In this work, when this case occurs for this algebra, the upper and lower bounds are switched and, consequently, a regular interval is created.

With this alteration, the new results of the centers and ranges for the different operations are presented in Table 3. The only difference when comparing them with Table 2 is that, now, the values of the ranges always correspond to an absolute value and, consequently, there can be no negative ranges. Even with this alteration, the Extended Interval Algebra still remains a vector space and the ranges can either decrease or increase their value with consecutive operations. However,

Table 3: Summarized results for the operations of the altered Extended Interval Algebra.

Operation	Center	Range
$y_1 + \beta$	$c_1 + \beta$	r_1
$y_1 + y_2$	$c_1 + c_2$	$r_1 + r_2$
$y_1 - y_2$	$c_1 - c_2$	$\lvert r_1 - r_2 \rvert$
βy_1	βc_1	$\lvert \beta r_1 \rvert$
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\lvert \beta_1 r_1 + \beta_2 r_2 \rvert$

when this algebra was tested in SPCA with interval variables, it was observed that the resulting intervals would often degenerate into intervals with very small ranges, which would be useless when analysing and subsequently manipulating the results. In SPCA it is preferred that the resulting intervals have a large enough range that could be representative of a real interval object, instead of having results where the ranges of the intervals are so small that they degenerate into single values. Therefore, Moore's Interval Algebra is still a better option than the Extended Interval Algebra for this statistical method.

4. Histogram Algebra

Developing an algebra for histogram-valued variables is more complex than for the interval case. In this work, an algebra is formalized, taking into consideration the arithmetic operations with histograms proposed by Irpino and Verde in [5].

4.1. Arithmetic operations with quantile functions

The definition of quantile function for histograms in (6) is the basis for the operations with histograms defined in [5]. However, these functions cannot be used directly in the operations and a transformation of the histograms has to be made beforehand.

Considering that our objective is to perform arithmetic operations in k histograms x_j, with $j = \{1, .., k\}$, where the number of subintervals for the j-th unit is n_j, our goal is to build histograms with the same number n of subintervals and where each of these subintervals is associated to the same cumulative probability w_i defined in (7). In this way, it is easier to perform the operations with quantile functions. This procedure will henceforth be called *harmonization*.

The first step of the harmonization procedure is, by taking into account the cumulative probabilities w_i that are a part of the quantile functions of the histograms we want to do arithmetic operations with, to build a new sorted and non-repetitive group of w_is that gathers all the different w_is from all the histograms. This harmonized set is denoted by $\mathcal{W}^\star = \{w_0^\star, w_1^\star, ..., w_i^\star, ..., w_n^\star\}$ in our work. The size of $\mathcal{W}^\star$ is n, which corresponds to the number of non-repeated different w_is from the histograms that take part in the operations. As a result, the elements $w_i^\star$ from the newly created set $\mathcal{W}^\star$ have the following properties:

1. $w_0^\star = 0$.
2. $w_n^\star = 1$.
3. $w_i^\star \neq w_j^\star, \forall_{i \neq j}, i, j \in \{0, 1, 2, .., n\}$.
4. $w_{i+1}^\star > w_i^\star, \forall_i, i \in \{0, 1, 2, .., n-1\}$.
5. $max\{n_1, n_2, .., n_k\} \leq n \leq \sum_{j=1}^{k} n_j - 1$.

The next step in the harmonization process is to build new quantile functions for the histograms using the elements of the set $\mathcal{W}^\star$ as the new bounds of the branches of the function. To find these bounds, it is necessary to compute the value of each $w_i^\star$ in the original quantile function of the histogram, $Q_{x_j}(w_i^\star)$, for $i \in \{0, 1, .., n\}$. Hence, the newly formed harmonized histogram, using the new set of cumulative probabilities $\mathcal{W}^\star$ for a histogram x_j, is given by

$$x_j^\star = \{[Q_{x_j}(w_0^\star), Q_{x_j}(w_1^\star)[, p_1^\star; ...; [Q_{x_j}(w_{i-1}^\star), , Q_{x_j}(w_i^\star)[, p_i^\star; ...; [Q_{x_j}(w_{n-1}^\star), Q_{x_j}(w_n^\star)], p_n^\star\}. \quad (10)$$

One of the main issues of the harmonization procedure is that, with each consecutive division of the original subintervals, smaller and smaller subintervals are created. At some point, these subintervals could become so small that the information provided by them would be meaningless. However, if we are not dealing with extreme cases, the harmonization process is a fairly good option to perform mathematical operations with histogram variables.

After having performed the harmonization procedure described previously, it is easy to perform arithmetic operations with the harmonized histograms in (10), using their harmonized quantile functions.

Under these conditions, the sum of two histograms is done by simply adding their harmonized quantile functions. The resulting centers and ranges of each subinterval i from the sum of two histograms, $x_1 + x_2$, are

$$\begin{aligned} c_i &= c_{i1}^\star + c_{i2}^\star, \\ r_i &= r_{i1}^\star + r_{i2}^\star, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \quad (11)$$

To compute the sum of a histogram with a constant $\beta \in \mathbb{R}$, we just have to add β to each branch of the quantile function. Therefore, the centers and ranges for the subinterval i of the histogram, resulting from the operation $x_1 + \beta$, are given by

$$\begin{aligned} c_i &= c_{i1} + \beta, \\ r_i &= r_{i1}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \quad (12)$$

Another arithmetic operation that is important to define is the multiplication by a constant $\beta \in \mathbb{R}$.

It is necessary to be careful when specifying this operation. The most correct method to compute βx_1, when $\beta < 0$, is firstly to do the transformation $Q_{x_1}(1-p)$ on the quantile function and then directly multiply the resulting function by β. The transformation $-Q_{x_1}(1-p)$, when performed on the quantile function Q_{x_1} of a histogram x_1, generates a new histogram that is the symmetric of x_1 with respect to the y-axis.

Accordingly, the product of a histogram x_1 and a real number β creates a new histogram, whose centers and ranges of its i-th subinterval are given by the expressions:

- for $\beta > 0$:

$$\begin{aligned} c_i &= \beta c_{i1}, \\ r_i &= \beta r_{i1}, \quad \text{with } i \in \{1, .., n\}; \end{aligned} \tag{13}$$

- for $\beta < 0$:

$$\begin{aligned} c_i &= \beta c_{n_1-i+1\,1}, \\ r_i &= |\beta| r_{n_1-i+1\,1}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{14}$$

The next arithmetic operation that is important to define is the difference between two histograms, $x_1 - x_2$. The first step in this operation is to obtain $-x_2$ according to the transformation of its quantile function $-Q_{x_2}(1-p)$ described previously. Afterwards, x_1 and $-x_2$ are harmonized and, finally, a regular addition of their quantile functions is performed. It is important to remark that, when carrying out this operation, the harmonization must always be done on the inverse of x_2, $-x_2$, and thus generate $(-x_2)^\star$. Instead, if the harmonization is performed firstly on x_2, and $x_2^\star$ is obtained, followed by the calculation of its inverse $-(x_2^\star)$, an additional harmonization may need to be applied, which is always undesirable.

It is considered that the transformation $-Q_{x_j}(1-p)$ and subsequent harmonization with the other histogram participating in the operation, yields the histogram $(-x_j)^\star$. In our work, $(-x_j)^\star$ is represented with the notation of a set of centers and ranges for each of the subintervals, as follows:

$$(-x_j)^\star = \{(-c_{nj}^{\diamond\star}, r_{nj}^{\diamond\star}), p_{nj}^{\diamond\star}; ..; (-c_{1j}^{\diamond\star}, r_{1j}^{\diamond\star}), p_{1j}^{\diamond\star}\}, \tag{15}$$

where the symbol $\diamond$ before the $\star$ on top of the harmonized bounds, centers, ranges, and associated probabilities represents the fact that the original order of the subintervals was reversed firstly and only afterwards was the harmonization performed. Therefore, in the operation $x_1 - x_2$, the order of the subintervals is reversed in the histogram $(-x_2)^\star$. Hence, the centers and ranges for the resulting subinterval i of $x_1 - x_2$ are given by

$$\begin{aligned} c_i &= c_{i1}^\star - c_{n-i+1\,2}^{\diamond\star}, \\ r_i &= r_{i1}^\star + r_{n-i+1\,2}^{\diamond\star}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{16}$$

4.2. Formalization of the Histogram Algebra

The first objective of our work is to create an algebra that sums up all the knowledge from the previously discussed operations with the quantile functions of histograms. This is achieved by using the information contained in the expressions (11), (12), (13), (14) and (16), regarding the resulting centers and ranges for each subinterval i of the histogram generated for each operation. Table 4 summarizes these results.

Table 4: Centers and ranges of the subintervals resulting from the quantile operations.

Operation	Center i	Range i
$x_1 + x_2$	$c_{i1}^\star + c_{i2}^\star$	$r_{i1}^\star + r_{i2}^\star$
$x_1 + \beta$	$c_{i1} + \beta$	r_{i1}
βx_1, $\beta \geq 0$	βc_{i1}	βr_{i1}
βx_1, $\beta < 0$	$\beta c_{n_1-i+1\,1}$	$\lvert\beta\rvert r_{n_1-i+1\,1}$
$x_1 - x_2$	$c_{i1}^\star - c_{n-i+1\,2}^{\diamond\star}$	$r_{i1}^\star + r_{n-i+1\,2}^{\diamond\star}$

Comparing these results with those obtained in Table 1 for the operations with Moore's Interval Algebra, it is possible to infer that, for each subinterval i, the operations with quantile functions for histograms are along the lines of Moore's Interval Algebra. Accordingly, while the centers of the subintervals can take any in value in $\mathbb{R}$, the ranges are always non-negative and expand with each consecutive operation.

For the linear combination $\beta_1 x_1 + \beta_2 x_2$ there are four distinct cases, depending on the sign of the constants β_1 and β_2. Table 5 displays the results.

By analyzing Table 5, one can conclude that it would be difficult to merge these four cases into a single formula. This is mainly due to the inversion of the indices of the histogram subintervals that were multiplied by a negative weight β. To solve this issue, a special notation (with the symbol $\bullet$) is used, in order to merge the cases where $\beta_j < 0$ and $\beta_j > 0$. Under this new notation, $c_{ij}^\bullet$ and $r_{ij}^\bullet$ are defined by

$$c_{ij}^\bullet = \begin{cases} c_{ij}^\star, & \beta_j \geq 0 \\ -c_{n-i+1\,j}^{\diamond\star}, & \beta_j < 0 \end{cases}, \tag{17}$$

$$r_{ij}^\bullet = \begin{cases} r_{ij}^\star, & \beta_j \geq 0 \\ r_{n-i+1\,j}^{\diamond\star}, & \beta_j < 0 \end{cases}, \tag{18}$$

with $i \in \{1, .., n\}$.

As the probabilities also have their order reversed when $\beta_j < 0$, it is also necessary to define $p_{ij}^\bullet$:

$$p_{ij}^\bullet = \begin{cases} p_{ij}^\star, & \beta_j \geq 0 \\ p_{n-i+1\,j}^{\diamond\star}, & \beta_j < 0 \end{cases}. \tag{19}$$

Hence, using the expressions (17) and (18), the resulting centers and ranges of the subintervals of a linear combination of two histograms,

Table 5: Different cases for the linear combination of two quantile functions.

$\beta_1 x_1 + \beta_2 x_2$	Center i	Range i
$\beta_1 \geq 0, \beta_2 \geq 0$	$\beta_1 c^{\star}_{i1} + \beta_2 c^{\star}_{i2}$	$\beta_1 r^{\star}_{i1} + \beta_2 r^{\star}_{i2}$
$\beta_1 < 0, \beta_2 < 0$	$\lvert\beta_1\rvert(-c^{\diamond\star}_{n-i+1\,1}) + \lvert\beta_2\rvert(-c^{\diamond\star}_{n-i+1\,2})$	$\lvert\beta_1\rvert r^{\diamond\star}_{n-i+1\,1} + \lvert\beta_2\rvert r^{\diamond\star}_{n-i+1\,2}$
$\beta_1 < 0, \beta_2 \geq 0$	$\lvert\beta_1\rvert(-c^{\diamond\star}_{n-i+1\,1}) + \beta_2 c^{\star}_{i2}$	$\lvert\beta_1\rvert r^{\diamond\star}_{n-i+1\,1} + \beta_2 r^{\star}_{i2}$
$\beta_1 \geq 0, \beta_2 < 0$	$\beta_1 c^{\star}_{i1} + \lvert\beta_2\rvert(-c^{\diamond\star}_{n-i+1\,2})$	$\beta_1 r^{\star}_{i1} + \lvert\beta_2\rvert r^{\diamond\star}_{n-i+1\,2}$

$\beta_1 x_1 + \beta_2 x_2$, with β_1, $\beta_2 \in \mathbb{R}$, are

$$\begin{aligned} c_i &= |\beta_1| c^{\bullet}_{i1} + |\beta_2| c^{\bullet}_{i2}, \\ r_i &= |\beta_1| r^{\bullet}_{i1} + |\beta_2| r^{\bullet}_{i2}. \end{aligned} \tag{20}$$

By using this information, it is possible to deduce a general formula for the linear combination of histograms in this algebra. Firstly, it is necessary to define the matrices $\boldsymbol{C}^{\bullet}$ and $\boldsymbol{R}^{\bullet}$, with the previously defined $c^{\bullet}_{ij}$ and $r^{\bullet}_{ij}$, respectively. The rows of these two matrices are associated with the subinterval i, whereas the j-th column corresponds to the harmonized histogram j. It is also considered a vector $\boldsymbol{p}^{\bullet}_j$ of the n harmonized probabilities associated to each histogram j, with $\boldsymbol{p}^{\bullet}_j = (p^{\bullet}_1, .., p^{\bullet}_i, .., p^{\bullet}_n])^t$.

Hence, being $\boldsymbol{\beta}^t \boldsymbol{x}$ a linear combination of k histograms $\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$, with $\boldsymbol{\beta} = (\beta_1, .., \beta_k)^t$, and $\boldsymbol{x} = (x_1, .., x_k)^t$, its general expression is given by

$$\begin{aligned} \boldsymbol{\beta}^t \boldsymbol{x} = \{ \boldsymbol{C}^{\bullet}_1 |\boldsymbol{\beta}| \pm \frac{1}{2} \boldsymbol{R}^{\bullet}_1 |\boldsymbol{\beta}|, p^{\bullet}_1; \boldsymbol{C}^{\bullet}_2 |\boldsymbol{\beta}| \pm \frac{1}{2} \boldsymbol{R}^{\bullet}_2 |\boldsymbol{\beta}|, p^{\bullet}_2; ...; \\ ; \boldsymbol{C}^{\bullet}_n |\boldsymbol{\beta}| \pm \frac{1}{2} \boldsymbol{R}^{\bullet}_n |\boldsymbol{\beta}|, p^{\bullet}_n \}, \end{aligned} \tag{21}$$

where $\boldsymbol{C}^{\bullet}_i$ and $\boldsymbol{R}^{\bullet}_i$ are representative of the row i of the matrices $\boldsymbol{C}^{\bullet}$ and $\boldsymbol{R}^{\bullet}$, respectively.

This algebra fails two of the axioms of vector space: the existence of an additive inverse, and also the distributivity of scalar multiplication with respect to field addition. Nonetheless, the remaining axioms hold and are useful, as is the case of the associativity and commutativity of addition.

5. Symbolic Principal Components Analysis

The last goal of this work is to use the algebra for histogram variables, based on the operations with quantile functions, to find a way to improve the results obtained for different methods of SPCA applied to histograms.

5.1. Overview of PCA and SPCA methods

PCA is a statistical method that transforms the original variables of a random vector into a new set of variables by using a simple orthogonal transformation. The resulting new set of variables are called Principal Components (PCs) and they are uncorrelated with one another.

The values of the PCs are usually denominated as *component scores* and the weights by which the original variables were multiplied to reach these scores are called loadings. In the conventional framework, these *loadings* usually correspond to the eigenvectors of the covariance matrix of the original data. Thus, considering that our original data is represented by a $(k \times p)$ matrix $\boldsymbol{X}$ with p variables, whose columns are associated to the variables and the rows to the k observations, and that $\boldsymbol{\delta}_l$ is the l-th eigenvector of the covariance matrix of $\boldsymbol{X}$, with $l \in (1, .., p)$, the vector with the scores for the l-th Principal Component ($\boldsymbol{PC}_l$) is given by

$$\boldsymbol{PC}_l = \boldsymbol{X} \boldsymbol{\delta}_l. \tag{22}$$

In works where the SPCA is applied to histogram-valued variables, several different ways to compute the PC scores are used. Yet, in none of them are the PC scores defined as the linear combination of histograms, as proposed in this work.

It is important to analyze how the covariance(correlation) matrices are defined in some of the other works in this area, and also how the authors choose to represent the resulting principal components. The authors, articles and corresponding definition of the covariance(correlation) matrix for histograms, which are used here, are:

- M. Kallyth, E.Diday in [8] and M. Chen, H. Wang in [3]:
  In both works the definition of sample symbolic covariance used to build the covariance matrix is denoted as cov_1. It is given by

$$cov_1(X_1, X_2) = \frac{1}{k} \sum_{j=1}^{k} \bar{x}_{j1} \bar{x}_{j2} - \bar{x}_1 \bar{x}_2, \tag{23}$$

  where $\bar{x}_l = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{n_{jl}} c_{ijl} p_{ijl}$ and $\bar{x}_{jl} = \sum_{i_l=1}^{n_{jl}} c_{i_l jl} p_{i_l jl}$, $l = \{1, 2\}$.
  The sample covariance matrix obtained through this method is denoted by $\boldsymbol{\Sigma}^{(1)}$ in this work. The resulting PCs are represented either as hypercubes or as intervals lengths in [8] and with a probability density function in [3].

- J. Le-Rademacher and L.Billard in [7]:
  Here the definition of sample symbolic covariance used is denoted as cov_2 and is given by

$$\begin{aligned} cov_2(X_1, X_2) = \frac{1}{3k} \sum_{j=1}^{k} \sum_{i_1=1}^{n_{j1}} \sum_{i_2=1}^{n_{j2}} p_{i_1 j1} p_{i_2 j2} \times \\ \times G_{i_1 j1} G_{i_2 j2} [Q_{i_1 j1} Q_{i_2 j2}]^{\frac{1}{2}}, \end{aligned} \tag{24}$$

with $Q_{i_l jl} = (a_{i_l jl} - \bar{x}_l)^2 + (a_{i_l jl} - \bar{x}_l)(b_{i_l jl} - \bar{x}_l) + (b_{i_l jl} - \bar{x}_l)^2$ and

$$G_{i_l jl} = \begin{cases} -1, & \bar{x}_{jl} \leq \bar{x}_l \\ 1, & \bar{x}_{jl} > \bar{x}_l \end{cases},$$

where $l = \{1, 2\}$. This definition of sample covariance is used to build the covariance matrix, which is denoted by $\mathbf{\Sigma}^{(2)}$.

This method of SPCA is based on the geometric construction of polytopes using the subintervals of the histograms, which can then be converted back into histograms. This is the only work of a SPCA method, among the ones found, which represents the PCs as histograms. However, this polytope method can be somewhat complex and hard to interpret.

- M. Ichino in [4]:
  This SPCA method for histogram variables starts by computing quantiles from the quantile function of the histograms. For a sample of size k of a p-dimensional random vector $\boldsymbol{X} = (\boldsymbol{X}_1, .., \boldsymbol{X}_l, .., \boldsymbol{X}_p)^t$, a number d is chosen arbitrarily with $1 \leq d \leq k$. Afterwards, $d + 1$ quantiles, Q_{ijl}, are generated for each observation j of a variable X_l, x_{jl}, through the expression $Q_{ijl} = Q_{x_{jl}}(\frac{i}{d})$, with $i = \{0, 1, ..., d\}$, $j = \{1, ..., k\}$, $l = \{1, ..., p\}$.

  Then, a numerical matrix $\boldsymbol{Q}$ with $k \times (d+1)$ rows and p columns is constructed in such a way that the quantiles Q_{ijl} for the k observations of a variable X_l are introduced in a sequential order in the column l of the matrix $\boldsymbol{Q}$. A sample correlation matrix is computed next for the data table presented in $\boldsymbol{Q}$, using either the Spearman's or the Pearson's correlation coefficient. The sample correlation matrix used in our work, based on this method, is denoted by $\mathbf{\Sigma}^{(3)}$, and only the Spearman's correlation coefficient is used for its computation, as it gave better results in [4].

5.2. SPCA using linear combinations of histograms

To improve the current SPCA methods that deal with histogram-valued variables, it is shown in this work how to obtain histogram-valued scores. This is achieved by applying the previously defined algebra to compute the linear combinations of histograms.

Before proceeding, it is necessary to define some new matrices for the centers, ranges, and associated probabilities of the histograms, $\boldsymbol{C}_j$, $\boldsymbol{R}_j$, and $\boldsymbol{P}_j$. These matrices aggregate all the information of an observation j, for all the p variables of a data set. The rows of these matrices are associated to the subintervals i of the observation j, while the columns are associated to the variables X_l.

The harmonization procedure is then applied to all the j-th observations of the p histogram variables included in the matrices $\boldsymbol{C}_j$, $\boldsymbol{R}_j$, and $\boldsymbol{P}_j$, which enables us to make arithmetic operations with them. Therefore, by applying the expressions (17), (18), and (19), it is possible to build the harmonized matrices $\boldsymbol{C}_j^{\bullet}$, $\boldsymbol{R}_j^{\bullet}$, and $\boldsymbol{P}_j^{\bullet}$. Afterwards, the PCs are generated through the linear combination of the p variables of a data set for each of their k observations. Accordingly, by applying expression (21), each observation j of the l-th PC (score), PC_{jl} is obtained:

$$PC_{jl} = (\boldsymbol{C}_j^{\bullet} |\boldsymbol{\delta}_l| \pm \frac{1}{2} \boldsymbol{R}_j^{\bullet} |\boldsymbol{\delta}_l|, \boldsymbol{P}_{j1}^{\bullet}), \qquad (25)$$

where $\boldsymbol{P}_{j1}^{\bullet}$ relates to the vector corresponding to the first column of the matrix $\boldsymbol{P}_j$, whose columns are all equal. This process is repeated for the k observations of the data set, thus generating a $(k \times 1)$ vector with the l-th Principal Component scores, $\boldsymbol{PC}_l$, and for the p different eigenvectors $\boldsymbol{\delta}_l$, which generates p histogram-valued Principal Components $\boldsymbol{PC}_l$, as desired.

6. Analysis of data sets

The final step of this work consisted in implementing the expression (25) for the computation of the PCs of a real data set. The eigenvectors, $\boldsymbol{\delta}_l$, in that expression were obtained according to the several definitions of covariance or correlation matrices, $\mathbf{\Sigma}^{(1)}$, $\mathbf{\Sigma}^{(2)}$, and $\mathbf{\Sigma}^{(3)}$, defined previously. Two examples, based on real data sets were explored using SPCA: Iris flower [8] and Hardwood [1]. Based on the proposed definition of linear combinations of histograms, the associated scores were obtained. All the computations presented next were done with the use of the R software [10].

6.1. Iris data set

The first data set where this method was implemented is the histogram-valued Iris flower data set taken from [8]. This data set corresponds to a transformation of the conventional Iris flower data set, which originally consists of 150 observations belonging to three species of the iris flower (50 observations per species), characterized by four variables. By applying a K-means on this conventional data, according to [8], the 150 observations were classified into 10 groups. The information of the observations that belonged to each of these 10 groups was then merged to form 10 histogram units. The four histogram variables of this data describe the following features of each Iris flower:

- X_1 - sepal width;
- X_2 - petal width;
- X_3 - petal length;

- X_4 - sepal length.

By computing the sample symbolic covariance, according to definition cov_1 in (23), of this data set, the eignevectors, $\boldsymbol{\delta}^{(1)}$, of $\boldsymbol{\Sigma}^{(1)}$ obtained are

$$\boldsymbol{\delta}^{(1)} = \begin{bmatrix} 0.053 & 0.742 & -0.395 & 0.539 \\ -0.342 & -0.069 & -0.813 & -0.467 \\ -0.854 & -0.214 & 0.112 & 0.461 \\ -0.389 & 0.631 & 0.413 & -0.528 \end{bmatrix}.$$

The first PC obtained with this method explains 93.11% of the variance, the second 5.88%, the third 0.9%, and the fourth 0.097%. Through the observation of the first eigenvector, it is possible to infer that, for the definition of the first PC, the variable that has more weight (in a negative proportion) is, by far, X_3, which corresponds to the petal length, followed by X_4 and X_2, which have a similar weight and are related to the petal width and sepal length respectively. The variable X_1, which is associated to the sepal width, is practically irrelevant for the definition of $\boldsymbol{PC}_1$. Therefore, the lower the value of the first PC score is, the higher the value of the petal length will be and the higher the petal width and sepal length. For the second PC, the most relevant variables are X_1 and X_4, both related to the size of the sepal. As these variables affect the definition of $\boldsymbol{PC}_2$ in a positive way, the wider and larger the sepals of a set of flowers that formed a histogram observation are, the higher the score of that observation on the second PC is.

If we represent the PC scores observed through regular histogram graphs, it is hard to find patterns in the data. To overcome this difficulty, a joint probability analysis between $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$ was performed. To achieve this, a graph was constructed, where the observations from $\boldsymbol{PC}_1$ are represented along the x-axis and the units from $\boldsymbol{PC}_2$ along the y-axis. Rectangles are then built according to the bounds of the subintervals of the histogram units of $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$. The number inside each rectangle corresponds to the product of the probabilities associated to the subintervals which originated it. The higher that value is, in comparison with all the other probabilities of that observation, the darker the color from that rectangle will be. This representation is displayed in Figure 1. Through its analysis, it can be concluded that there is considerable variability in the distribution of the histograms among all the observations. However, it is now easier to find patterns in the data and reach some conclusions. Therefore, observation 5 seems to have been created by the aggregation of flowers with a medium petal and sepal width and length; observation 8 for the flowers also with a medium sepal width and length, but with a lower value for the petal length; observation 3 from the irises with a medium sepal size, but a high petal length; observation 9 from the flowers with a large sepal size and a wide range for the petal length; observation 6 from the ones with low values of both sepal and petal size; observation 7 from the irises with high values for both the sepal and petal size; and observations 1, 2, 4 and 10 from flowers with a wide range of values for both the petal and sepal size.

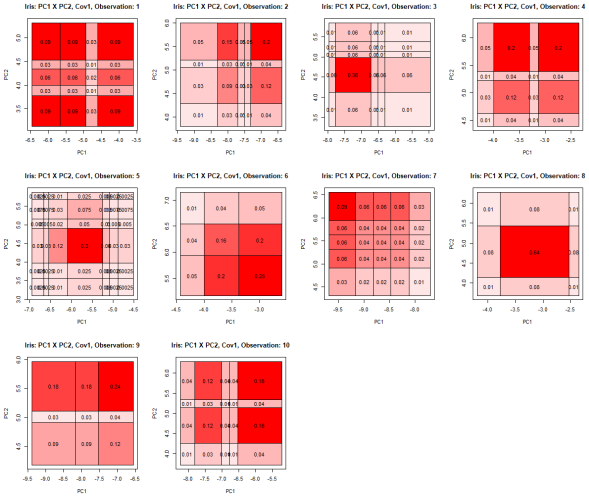

Figure 1: Joint probability of PC_1 and PC_2 scores when cov_1 is applied to the Iris data set.

It is also interesting to check what the sample covariance matrix for the PCs is, when the definition of covariance cov_1 in (23) is used (computation based on the obtained scores). This matrix, represented as $\boldsymbol{\Sigma}^{(1)}_{PC}$, is the following:

$$\boldsymbol{\Sigma}^{(1)}_{PC} = \begin{bmatrix} 3.782 & 0 & 0 & 0 \\ 0 & 0.239 & 0 & 0 \\ 0 & 0 & 0.037 & 0 \\ 0 & 0 & 0 & 0.004 \end{bmatrix}.$$

The values of the main diagonal of $\boldsymbol{\Sigma}^{(1)}_{PC}$ correspond to the variance explained by each PC and they match the percentages of total variance presented previously. The remaining values are equal to zero, which is to be expected, since the PCs are supposed to be uncorrelated. Hence, everything seems to match the expected results for this case.

When the definition of sample covariance cov_2 is used, we have that the first PC explains 92.04% of the variance, the second explains 6.69%, the third 0.96% and the fourth 0.35%. The first two eigenvectors from $\boldsymbol{\delta}^{(1)}$ and $\boldsymbol{\delta}^{(2)}$ obtained are very similar, which means that the first and second PCs obtained for both cases have a similar interpretation and lead to similar scores. Accordingly, the joint probability graph of $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$ for cov_2 is very similar to the one obtained for cov_1 in Figure 1, and the same conclusions can be reached. For these reasons, the inclusion of $\boldsymbol{\delta}^{(2)}$ and the joint probability graph were omitted from this work.

By computing the sample covariance matrix for the PCs, reached through cov_2 in (24), $\mathbf{\Sigma}_{PC}^{(2)}$, we obtain:

$$\mathbf{\Sigma}_{PC}^{(2)} = \begin{bmatrix} 3.914 & -0.236 & -0.136 & 0.358 \\ -0.236 & 0.431 & -0.015 & 0.159 \\ 0.136 & -0.015 & 0.230 & 0.098 \\ 0.358 & 0.159 & 0.098 & 0.327 \end{bmatrix}.$$

In $\mathbf{\Sigma}_{PC}^{(2)}$, the values of the main diagonal do not correspond to the percentages of the explained variance for the PCs calculated through the eigenvalues of $\mathbf{\Sigma}^{(2)}$. This may be explained by the fact that $cov_2(X, X)$ is not considered as a regular variance value. It is also possible to observe that the remaining values of $\mathbf{\Sigma}_{PC}^{(2)}$ are not equal to zero, as it was expected to happen due to the supposed orthogonality condition that in the conventional case leads to the construction of uncorrelated PCs.

Finally, for the correlation matrix, $\mathbf{\Sigma}^{(3)}$, the corresponding eigenvectors $\boldsymbol{\delta}^{(3)}$, obtained by considering $d = 4$ quantiles and the Spearman's correlation coefficient, are

$$\boldsymbol{\delta}^{(3)} = \begin{bmatrix} -0.184 & 0.944 & -0.219 & 0.164 \\ -0.565 & -0.188 & -0.669 & -0.444 \\ -0.567 & -0.242 & 0.025 & 0.787 \\ -0.570 & 0.122 & 0.709 & -0.395 \end{bmatrix}.$$

In this case, all the columns of $\boldsymbol{\delta}^{(3)}$ are significantly different, when comparing them with the two previous cases. Here, we obtain a first PC that explains 71.84% of the total variance of the data, an amount which is much lower than the ones obtained for the first PC in the two previous methods. Despite the differences in the eignenvectors, the distribution of the histograms continues to be very similar when performing a graphic analysis. Consequently, this graph has also been omitted.

The sample correlation matrix for the PCs, $\mathbf{\Sigma}_{PC}^{(3)}$, obtained through the same method, is

$$\mathbf{\Sigma}_{PC}^{(3)} = \begin{bmatrix} 1 & 0.778 & 0.237 & -0.189 \\ 0.778 & 1 & 0.608 & 0.129 \\ 0.237 & 0.608 & 1 & 0.786 \\ -0.189 & 0.129 & 0.786 & 1 \end{bmatrix}.$$

In the main diagonal of this matrix all the values are equal to 1, as it is always expected to happen for a correlation matrix, since $corr(X, X) = 1$. However, as it also happened for the cov_2 case, the remaining values of this matrix are not equal to zero. This highlights the fact that this method does not lead to uncorrelated PCs, according to definition $\mathbf{\Sigma}^{(3)}$.

6.2. Hardwood data set

The method developed in this work was applied to a second data set, obtained from a US Geological Survey of Hardwood Trees that can be found in [1]. This data set has 16 objects, which correspond to different species of hardwood trees, each being described by the following eight features related to the regions where the species can be found:

- X_1: Annual temperature (ºC);
- X_2: January temperature (ºC);
- X_3: July temperature (ºC);
- X_4: Annual precipitation (mm);
- X_5: January precipitation (mm);
- X_6: July precipitation (mm);
- X_7: Growing degree days on $5ºC \times 1000$;
- X_8: Moisture index.

Before applying the SPCA, all these variables were standardized, according to appropriate definitions of sample symbolic mean and variance.

When cov_1 is used to compute the sample covariance matrix of this data, the first two eigenvectors of $\boldsymbol{\delta}^{(1)}$ are

$$\begin{bmatrix} -0.4700 & 0.0390 \\ -0.4352 & 0.1620 \\ -0.4515 & -0.1578 \\ 0.0194 & -0.6262 \\ 0.2640 & -0.1098 \\ -0.2922 & -0.3232 \\ -0.4800 & 0.0741 \\ 0.0002 & -0.6581 \end{bmatrix}$$

The first PC obtained can be interpreted as the weighted sum of the variables X_1, X_2, and X_3, which are related to temperature, and X_7, related to a measure of heat accumulation. For the second PC, the two most significant variables for its definition are X_4, related to annual precipitation, and X_8, related to moisture.

The graph of the joint probabilities of the PCs obtained is displayed in Figure 2. It can be observed that there seem to be increasingly higher probabilities the closer you are to the center. This may be a consequence of the standardization of the data, which was performed beforehand. In cases like this, it is hard to reach any relevant conclusions through the analysis of the joint probability graph of the first and second PCs.

For all the three different definitions of covariance/correlation matrix, the results obtained in this data set showed similar distributions of the histograms in the joint probability graphs of $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$, as it happened with the Iris data set. Moreover, an uncorrelated structure of the PCs also occurred only when the definition of sample covariance used was cov_1 in (23).

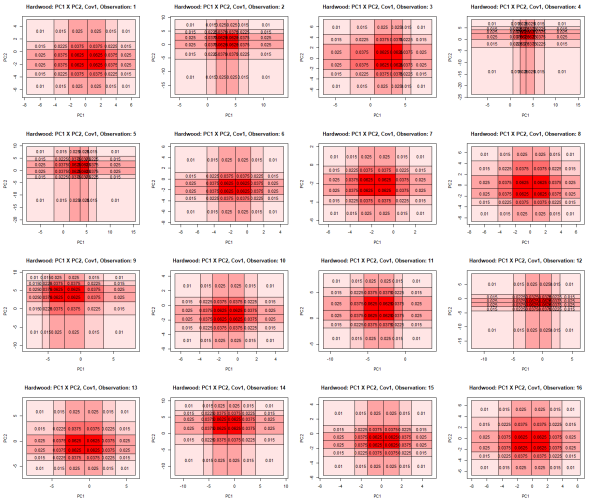

Figure 2: Joint probability between PC_1 and PC_2 scores when cov_1 is applied to the Hardwood data set.

7. Conclusions

In this work, we were able to find the general expression (21) for the computation of linear combinations, according to the histogram algebra with quantile function. This algebra follows a reasoning similar to that of Moore's Interval Algebra, namely the fact that the resulting ranges of the intervals/subintervals expand with each consecutive operation. This can be a disadvantage because when intervals/subintervals become very large, they may lose their significance. Nonetheless, it is still the best option to perform arithmetic operations with intervals and histograms.

This algebra was then applied to the SPCA statistical method to help build PCs, which are also histogram-valued variables, unlike what happens in the vast majority of the other works in the area. This was accomplished with the linear combination of histogram variables using the expression (25).

The proposed method was tested for three different definitions of covariance. The results obtained were very similar for the three definitions of covariance/correlation matrix used for both data sets, in terms of the distribution of the resulting histograms. However, only through the use of the definition of covariance cov_1 in (23), was it possible to verify that the Principal Components obtained were in fact uncorrelated, as well as to relate the main diagonal of the covariance matrix of the PCs with the values of the variance obtained for each PC. Therefore, the use of a covariance matrix computed from the definition of covariance cov_1 may lead to more trustworthy and easier to interpret results.

The joint probability graph that was used allowed us to find some patterns in the first data set. However, no relevant conclusions were reached for the second one. Therefore, even if this method can be useful in some data sets, other statistical methods also need to be applied for a deeper analysis. Accordingly, an advantage of using this method is that the resulting histogram-valued PCs that explain the majority of the variance in the data can be used to reduce the dimension of the data set. This reduced data set can afterwards be applied to improve symbolic classification methods, for example. Furthermore, the PCs obtained are uncorrelated symbolic variables, which can be helpful for many methods.

References

[1] Histogram data by the U.S. Geological Survey, Climate-Vegetation Atlas of North America. http://pubs.usgs.gov/pp/ p1650-b/. Last accessed December 16, 2019.

[2] H.-H. Bock and E. Diday. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, 2000.

[3] M. Chen, H. Wang, and Z. Qin. Principal component analysis for probabilistic symbolic data: a more generic and accurate algorithm. *Adv. Data Anal. Classif.*, 9:59–79, 2015.

[4] M. Ichino. The Quantile Method for Symbolic Principal Component Analysis. *Adv. Data Anal. Classif.*, 4(2):184–198, 2011.

[5] A. Irpino and R. Verde. A new Wassertein based distance for the hierarchical clustering of histogram symbolic data. In *Data Science and Classification. Proceedings of the 10th Conference of the International Federation of Classification Societies (IFCS'06)*, pages 185–192. Springer, 2006.

[6] E. Kaucher. Interval Analysis in the Extended Interval Space $\mathbb{R}$. *Fundamentals of Numerical Computation (Computer-Oriented Numerical Analysis). Computing Supplementum*, 2:33–49, 1980.

[7] J. Le-Rademacher and L. Billard. Principal component analysis for histogram-valued data. *Adv. Data Anal. Classif.*, 11:327–351, 2017.

[8] S. Makosso-Kallyth and E. Diday. Adaptation of interval PCA to symbolic histogram variables. *Adv. Data Anal. Classif.*, 6:147–159, 2012.

[9] R. Moore. *Interval Analysis*. Prentice-Hall, 1966.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.