TÉCNICO LISBOA

Histogram Algebra: an Application to Histogram Principal Component Analysis

Eduardo Manuel Jubilot Mendes

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisors: Professor Maria do Rosário de Oliveira Silva
Professor Lina Maria Mateus de Oliveira

Examination Committee

Chairperson: Prof. António Manuel Pacheco Pires
Supervisor: Prof. Maria do Rosário de Oliveira Silva
Member of the Committee: Prof. Sónia Manuela Mendes Dias

December 2019

Acknowledgments

First, I would like to thank both my supervisors: Professors Maria do Rosário de Oliveira Silva and Lina Maria Mateus de Oliveira. Their help, guidance and suggestions were invaluable for the development of this work.

I also thank Professor Sónia Dias and Margarida Vilela for kindly providing me with material when requested, which was essential to complete this work.

Finally, I would like to thank my parents for their support over the years.

Resumo

Nos últimos anos, a Análise de Dados Simbólicos tem vindo a tornar-se uma área cada vez mais importante da Estatísitica, devido à crescente complexidade dos dados a serem tratados. Um dos tipos de variáveis simbólicas mais utilizado é o histograma, que contém informação sobre a distribuição de probabilidades dos indíviduos que a originaram. Por este motivo, é importante encontrar formas de manipular e fazer operações aritméticas de um modo fácil com este tipo de variável. Dado que as variáveis intervalares constituem um tipo específico de variáveis histograma, estas merecerão também particular atenção neste trabalho.

Após ter incidido o foco na criação de uma álgebra de histogramas, baseada nas operações aritméticas com funções quantil, que é uma das formas possíveis de representar histogramas, é obtida uma expressão geral para o cálculo de combinações lineares entre histogramas.

A expressão mencionada acima é posteriormente utilizada para sugerir um novo método de estimação para a Análise Simbólica de Componentes Principais, quando este é aplicado a dados histograma. Como resultado, o método representa as observações originais, em forma de histograma multivariado, no espaço gerado pelas componentes principais simbólicas, ao invés do que sucede na larga maioria dos trabalhos nesta área. Este método é aplicado a dois conjuntos de dados, sendo depois analisadas as suas vantagens e limitações, tendo em conta os resultados obtidos.

Palavras-chave: Análise de Dados Simbólicos, variáveis histograma, funções quantil, álgebra de histogramas, covariância simbólica, Componentes Principais Simbólicas

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

Contents

List of Tables

List of Figures

Acronyms

cdf cumulative distribution function.

PC Principal Component.

PCA Principal Components Analysis.

SDA Symbolic Data Analysis.

SPCA Symbolic Principal Components Analysis.

Chapter 1

Introduction

1.1 Motivation

The ever-growing importance of Symbolic Data Analysis (SDA), 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 [1] 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.

1.2 Symbolic Data Analysis

In conventional data analysis each object is always characterized by a random vector composed by single numerical values or categories, related to each of the object variables. Therefore, considering that we have k objects from a conventional data set under these conditions, with p variables each, it is

possible to represent this data using an $(k \times p)$ matrix according to the following structure:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kp} \end{bmatrix},$$

where x_{ij} corresponds to the value the object i takes for the variable j, with $i \in \{1, 2, .., k\}$, and $j \in \{1, 2, .., p\}$. The values each of these x_{ij} take can either be a real number or a category depending on whether the j-th variable is numerical or categorical.

To illustrate this, consider a conventional data set, as the one represented originally in Table 1.1, related to health measures (heart rate and mean arterial pressure) along with age and gender (categorical variable that takes the value 1 when the patient is male and 0 when the patient is female) of five patients from a hospital. We can represent all this information in a matrix **X**, which is displayed in Figure 1.1. Since, we have observations from five different patients, each one characterized by four variables, the matrix **X** has five rows and four columns.

Table 1.1: Conventional data table with information from patients of a hospital.

Patient	Age	Gender	Heart rate (beats/min)	Mean arterial pressure (mm Hg)
1	40	M (1)	110	66
2	58	M (1)	89	100
3	39	F (0)	100	83
4	60	M (1)	95	79
5	50	F (0)	90	97

$$\mathbf{X} = \begin{bmatrix} 40 & 1 & 110 & 66 \\ 58 & 1 & 89 & 100 \\ 39 & 0 & 100 & 83 \\ 60 & 1 & 95 & 79 \\ 50 & 0 & 90 & 97 \end{bmatrix}$$

Figure 1.1: Matrix representation of the conventional data represented in Table 1.1.

As it was demonstrated with the previous example, conventional data can be easily converted into mathematical information, which can afterwards be analyzed using statistical methods. However, assigning only one single value to each observation of a variable might not be the most accurate way to represent it, since we would not be accounting for any variability or uncertainty in the data. That is where Symbolic Data Analysis (SDA) comes in. SDA includes the internal variation of the variables by representing data in complex data tables, where each cell can now contain a finite set of numerical values, categories, intervals or distributions, instead of just one single value/category, as it happened in

the conventional case.

There are different types of symbolic variables, depending on whether the variables are categorical or numerical and on the way that their variation is represented. If the realization of a random variable is a finite set of values (numerical or categorical), then that variable is denominated as a multi-valued variable. If each observation is a finite set of categories or real numbers, each one associated with a probability or frequency, then it is designated as a modal-valued variable. In SDA, a variable that takes an individual value, as it happens in the conventional data analysis, is a special case of a multi-valued-variable. This type of variables are called single-valued variables and can be considered in SDA. In the case where the observations of a random variable are intervals of real numbers, that variable is denominated as an interval variable. All these different types of symbolic variables can be observed in Table 1.2. In this table, the variable Age is a numerical modal-valued variable, Number of scored goals is a numerical multi-valued variable, and Height is an interval variable.

Table 1.2: Symbolic data table with information from the players of four football teams.

Team	Age	Number of scored goals	Height (cm)
A	$\{[17, 21[, 0.3; [21, 26], 0.7\}$	$\{1, 3, 5\}$	$[171, 198]$
B	$\{[18, 23[, 0.7; [23, 29], 0.3\}$	3	$[165, 186]$
C	$\{[16, 19[, 0.2; [19, 25[, 0.4; [25, 30], 0.4\}$	$\{3, 4\}$	$[178, 200]$
D	$\{[19, 23[, 0.6; [23, 25], 0.4\}$	$\{1, 2, 3, 5\}$	$[170, 188]$

The objects used in the SDA may correspond either to single individuals (micro-data) or to a collection of micro-data whose information was aggregated to create more complex symbolic objects (macro-data). If we are dealing with micro-data in SDA, the variability is intrinsic to each object, while for the macro-data the variability can also be brought up by the aggregation of the information of all the single units that form it. For instance, if we consider all the patients from the data set represented previously in Table 1.1 (micro-data) as a single object belonging to a Hospital A, it is possible to create a new symbolic object (macro-data) by aggregating all the information of these five patients. This generates new symbolic variables depending on the way that aggregation is made. For the three quantitative variables (Age, Heart rate, Mean arterial pressure), we can create three interval-valued variables by taking the minimum and maximum value for each variable among all of the five patients. As for the categorical variable associated with gender, we know that three of the patients are male and two are female. Therefore, we can create a modal-valued variable containing the probabilities of a patient from this sample being male or female. A new symbolic data set can created by repeating this procedure for the micro-data from other hospitals, as it is displayed in Table 1.3. The first object (Hospital A) from this new symbolic data set is a possible way to represent the macro-data originated from the micro-data presented in Table 1.1 using the method previously discussed.

Table 1.3: Symbolic data table with information from patients of several hospitals.

Hospital	Age	Gender	Heart rate (beats/min)	Mean arterial pressure (mm Hg)
A	$[40, 60]$	$\{M, \frac{3}{5}; F, \frac{2}{5}\}$	$[89, 110]$	$[66, 100]$
B	$[35, 55]$	$\{M, \frac{2}{3}; F, \frac{1}{3}\}$	$[85, 115]$	$[60, 110]$
C	$[42, 61]$	$\{M, \frac{1}{2}; F, \frac{1}{2}\}$	$[98, 130]$	$[54, 90]$
D	$[39, 76]$	$\{M, \frac{2}{5}; F, \frac{3}{5}\}$	$[67, 105]$	$[67, 108]$

When dealing with interval-valued variables, it is frequent to consider that, within this interval, micro-data follows a continuous uniform distribution, which models the ignorance about the true distribution of the micro-data, which most of the times is not observed. By splitting each of these intervals into several subintervals and associating a probability or weight to each subinterval, we obtain what is called a histogram-valued variable. The values inside each subinterval of these histograms are assumed to be uniformly distributed. As referred by Brito in [2], for different observations, the number and length of the histograms' subintervals may be different. In Table 1.4 we consider the interval-valued variables related to the Heart rate and Mean arterial pressure from Table 1.3 and represent them as histogram-valued variables, where each original interval was split into two subintervals and the corresponding probability was associated to each of them.

Table 1.4: Symbolic data table with histogram-valued variables.

Hospital	Heart rate (beats/min)	Mean arterial pressure (mm Hg)
A	$\{[90, 105[, 0.8; [105, 110], 0.2\}$	$\{[66, 85[, 0.6; [85, 100], 0.4\}$
B	$\{[85, 95[, 0.3; [95, 115], 0.7\}$	$\{[60, 90[, 0.5; [90, 110], 0.5\}$
C	$\{[98, 110[, 0.7; [110, 130], 0.3\}$	$\{[54, 70[, 0.6; [70, 90], 0.4\}$
D	$\{[67, 90[, 0.1; [90, 105], 0.9\}$	$\{[67, 87[, 0.5; [87, 108], 0.5\}$

SDA can be useful, for example, when the size of the sample is too large and we want to reduce the number of observations. This can be easily achieved by aggregating the information from a few single observations and transforming them into symbolic variables. The interval-valued and histogram-valued variables are a particularly useful way to represent the aggregated information. To construct interval variables, the information regarding the absolute minimum and maximum from a set of micro-data is used. However, in some practical cases, to overcome outliers, high (low) order quantiles are used instead of the maximum (minimum) observed values. The histogram variables are more complex and are obtained by splitting an initial interval into a group of subintervals with a probability (or frequency) associated to them. In this way, we have more information about the micro-data distribution of the data than if interval-valued variables are used. Both of these types of symbolic variables are the main focus of this thesis and are further analyzed in the following chapters.

1.3 Organization of the thesis

This work is divided into five chapters, besides the Introduction:

In Chapter 2, some general notions and definitions are presented for both interval-valued and histogram-valued variables, followed by an enumeration of several ways to represent these variables, including the representation through a quantile function, which is essential in the following chapters.

Chapter 3 starts with a summary of two algebras for interval variables: Moore's Interval Algebra and Extended Interval Algebra. Next, it is shown how to perform different arithmetic operations with quantile functions. These operations are then used to create an algebra for histograms and to find a general expression for the linear combination of histograms.

In Chapter 4, some definitions of symbolic descriptive statistical measures are presented: sample symbolic mean, variance, covariance, and correlation. This is followed by an example to check if these are good estimators.

As for Chapter 5, the previously defined histogram algebra, based on the quantile functions, is used to propose a new estimation method for the SPCA method. Accordingly, a general expression to obtain histogram-valued Principal Components through the linear combination of histograms is obtained, which is then applied to two data sets.

Finally, in Chapter 6 the conclusions are summarized and possible ideas for future work are presented.

Chapter 2

Interval and Histogram Data

The main focus of this work is related to interval and histogram-valued variables. Both are very useful ways to represent data that can be used in SDA. Firstly, it is important to present some general notions and definitions concerning each of them and how they can be formally defined by using an appropriate notation. This is followed by an enumeration of the several ways of representing these variables, which are used throughout the remainder of this work.

2.1 Definitions

Intervals can be seen as simpler versions of histograms and they are a very common way to represent information. There are different types of intervals, including discrete and continuous ones. However, this work only deals with the continuous intervals that are designated as real intervals. A real interval is a set of the real numbers that are in between two numbers, the bounds of the interval. For instance, the set of numbers that meet the condition $1 \leq x \leq 5$ correspond to an interval which contains all the real numbers between 1 and 5. Under these conditions, an interval can be defined in the following way:

$$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}, \text{ with } a, b \in \mathbb{R},$$

where a is the lower bound of the interval and b corresponds to the upper bound. These bounds can be closed or open, depending on whether the bounds themselves belong or not to the interval. For instance, an interval whose upper bound is open is defined as

$$[a, b[= \{x \in \mathbb{R} : a \leq x < b\}, \text{ with } a, b \in \mathbb{R}.$$

Hence, by representing data in an interval form, we only have information regarding two values of an object. In practical applications, these two values correspond to very high and low quantiles (for example the 1% and 99% quantiles), in order to overcome the existence of outliers in the micro-data. However, sometimes, having information about only two values of an object is not enough and more complex ways to represent these objects are necessary. Such is the case of histograms.

Histograms are more commonly used as a graphic way to represent the distribution of data. They consist of an x-axis, an y-axis and bars with different widths and heights along these axes. The width of

the bars of a histogram corresponds to their range of values in the x-axis, while the height of the bars on the y-axis and, in some cases, the area of the bar, show how frequently the corresponding range of values on the x-axis occurs in the data.

A histogram can be created from a general interval, which accounts for the whole range of values that the data being analyzed can take, by splitting that interval into smaller consecutive disjoint subintervals, which are called bins. For instance, in Figure 2.1, an initial interval $[1, 8]$ was divided into three smaller bins: $[1, 3[$, $[3, 5[$, and $[5, 8]$, to create a histogram. It is then counted how many observations of the original data fall into each of the bins, which give us the height of each bar along the y-axis. In the histogram represented in Figure 2.1, we have a total of 100 observations, 10 of which have a value between 1 and 3, 60 with value between 3 and 5, and the remaining 30 are positioned between 5 and 8.

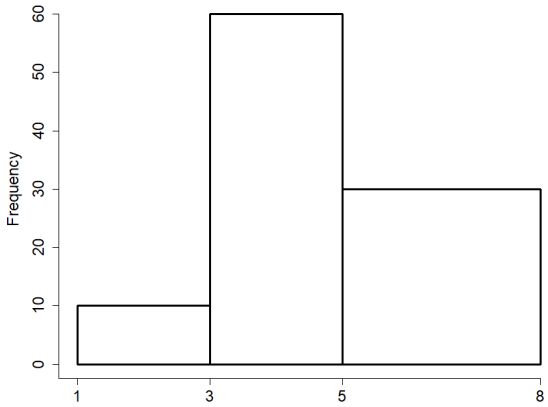


Figure 2.1: Example of a histogram with frequencies.

Instead of using the absolute frequencies in the y-axis of a histogram, it is also possible to use the relative frequencies (probabilities) of an observation falling into each of the bins. For example, Figure 2.2 displays the version of the histogram presented in Figure 2.1 using probabilities instead of frequencies. Each probability is obtained by simply dividing, for each bin, the corresponding frequency by the total number of observations, which in this case is 100. Since we are dealing with probabilities, the sum of the heights of all the bars of a histogram defined this way is always equal to 1. Throughout this work, probabilities are used instead of absolute frequencies for the visualization of the histograms, as they are easier to interpret.

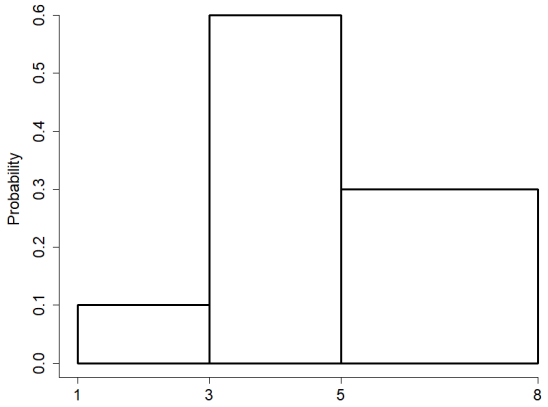


Figure 2.2: Example of a histogram with probabilities.

A visual analysis of a histogram allows us to observe how the data is distributed, in which ranges of values more observations can be found and also check for outliers. Due to these properties, histograms are considered one of the seven basic tools of quality, which are graphic methods used in quality control and were defined by K. Ishikawa in [3].

Having given these general notions and definitions about intervals and histograms, it is also important to find a more formal way to better define them, which can be achieved through the use of the symbolic variable notation.

First, it is necessary to define what a symbolic variable is. The definition that follows was taken from the Chapter 3 of the book of Bock and Diday [4].

Definition 2.1.1. *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 either micro-data or 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. Hence, it is easy to represent a sample of size k from a random vector of symbolic variables $(X_1, .., X_p)^t$ of size p, using this notation (see Table 2.1). Each column is, therefore, linked to one of the l symbolic variables, while the rows of the data table are associated to the corresponding observation j.

Table 2.1: General symbolic data table according to Definition 2.1.1.

	X_1	X_2	...	X_l	...	X_p
1	ϵ_{11}	ϵ_{12}	...	ϵ_{1l}	...	ϵ_{1p}
2	ϵ_{21}	ϵ_{22}	...	ϵ_{2l}	...	ϵ_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
j	ϵ_{j1}	ϵ_{j2}	...	ϵ_{jl}	...	ϵ_{jp}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	ϵ_{k1}	ϵ_{k2}	...	ϵ_{kl}	...	ϵ_{kp}

For example, if we are dealing with interval-valued variables, we have

$$\epsilon_{jl} = [a_{jl}, b_{jl}] \subset \mathbb{R}, \text{ with } a_{jl} \leq b_{jl}.$$

As for the histogram-valued variables, they are a particular case of modal-valued variables. These are a type of symbolic variables that occur when the set B is a set of distributions. One of the most used forms to describe the j-th observation of modal-valued variable X, x_j, is the following:

$$x_j = \{y_{ij}, p_{ij}; i \in \{1, ..., n_j\}\}.$$

In the previous definition, each y_{ij} is an individual belonging to a categorical/quantitative symbolic variable, and p_{ij} is a measure associated to y_{ij} (for histogram variables, it is considered that each p_{ij} corresponds to a probability and y_{ij} to a subinterval or bin).

For a modal-valued variable to be considered a histogram-valued variable, certain conditions have to be met:

1. The $y_{ij}, i \in \{1, ..., n_j\}$ must correspond to ordered and disjoint intervals. Each y_{ij} corresponds to a subinterval. For each x_j there can be multiple subintervals y_{ij}. Their number is equal to n_j.

2. $0 \leq p_{ij} \leq 1$, is the probability associated with each subinterval y_{ij}.

3. $\sum_{i=1}^{n_j} p_{ij} = 1$, *i.e.*, the sum of the probabilities for all subintervals, y_{ij}, is equal to 1.

Thus, for histogram-valued variables, we have the following definition:

Definition 2.1.2. *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.*

Commonly, among the symbolic community, a histogram x_j is usually written as

$$x_j = \{y_{1j}, p_{1j}; y_{2j}, p_{2j}; ...; y_{ij}, p_{ij}; ...; y_{n_j j}, p_{n_j j}\}.$$

Taking into consideration this previous definition, interval-valued variables can be seen as a particular case of a histogram-valued variable where $n_j = 1$. Hence, we have only one interval with an associated probability of 1, and thus:

$$x_j = y_j = [a_j, b_j].$$

It is considered that the micro-data associated with an interval or subinterval follows a uniform distribution. This assumption represents the ignorance about the true distribution of the micro-data within a subinterval.

It is also important to note that, as in Definition 2.1.2., the histogram-valued variables are represented by X, while the interval-valued variables are represented by Y, throughout this work. On the other hand, the sets of real histogram and interval objects are represented by $\mathcal{X}$ and $\mathcal{Y}$, respectively.

2.2 Representation

There are several different ways to represent intervals and histograms. Previously, it was seen that intervals are generally represented as follows:

$$y = [a, b] \quad \text{with } a, b \in \mathbb{R};\ a \leq b, \tag{2.1}$$

where a is the lower bound and b the upper bound of the interval y. It was also observed that histograms are characterized by a set of subintervals, such that the i-th subinterval of x_j is represented by $[a_{ij}, b_{ij}[$, with the associated probability p_{ij}. Admitting this notation, a histogram x_j with n_j subintervals can be represented as

$$x_j = \{[a_{1j}, b_{1j}[, p_{1j}; [a_{2j}, b_{2j}[, p_{2j}; ...; [a_{ij}, b_{ij}[, p_{ij}; ...; [a_{n_jj}, b_{n_jj}], p_{n_jj}\}, \tag{2.2}$$

with $a_{ij}, b_{ij} \in \mathbb{R}$; $a_{ij} \leq b_{ij}$; $a_{i+1j} = b_{ij}$; $0 \leq p_{ij} \leq 1$; $\sum_{i=1}^{n_j} p_{ij} = 1$; $i \in \{1, .., n_j\}$.
Example 2.2.1 illustrates the corresponding graphic representation of a histogram represented according to the notation in (2.2).

Example 2.2.1. *Considering that we have a histogram x represented using the structure in (2.2) as follows:*

$$x = \{[0, 3[, 0.1; [3, 4[, 0.3; [4, 8[, 0.4; [8, 15], 0.2\}.$$

This matches a histogram where the interval $[0, 15]$ was split into four subintervals: $[0, 3[$, $[3, 4[$, $[4, 8[$, and $[8, 15]$, with associated probabilities of 0.1, 0.3, 0.4 and 0.2, respectively, whose sum is equal to one. The graphic representation of this histogram can be seen in Figure 2.3.

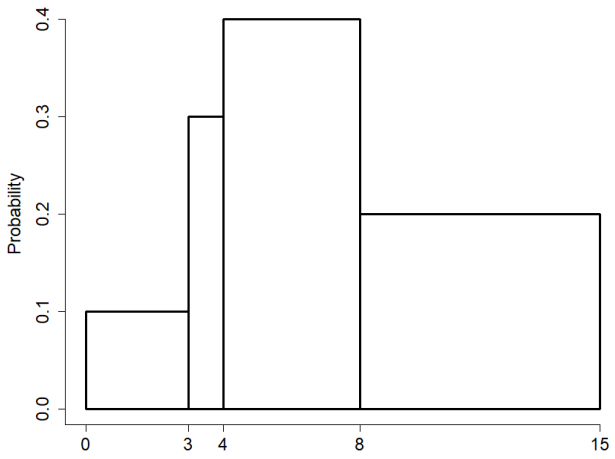


Figure 2.3: Histogram x associated with the Example 2.2.1.

□

However, using the bounds of the interval/subinterval is sometimes not the best way to represent and manipulate this type of data. An alternative is to use the center and range of the interval/subinterval instead. The center (c_y) of an interval $y = [a, b]$ is the middle point between its two bounds and is given by the expression

$$c_y = \frac{a + b}{2}. \tag{2.3}$$

The range (r_y) of an interval $y = [a, b]$ corresponds to its length, which is obtained from the distance between its two bounds:

$$r_y = b - a. \tag{2.4}$$

Since $b \geq a$ is always a condition for an interval, the range is always a non-negative value. The same does not happen for the center of a real interval, which can take any value in $\mathbb{R}$. An interval where $b = a$ has a center equal to a and b, and range zero, which corresponds to the single-valued case ($[a, a] = a$).

Using the previous definitions of center and range of an interval, it is possible to rewrite the expression (2.1) for the bounds of an interval as

$$y = [c_y - \frac{r_y}{2}, c_y + \frac{r_y}{2}] \quad \text{with } c_y \in \mathbb{R}, r_y \in \mathbb{R}_0^+. \tag{2.5}$$

In a more simplified manner, it is also possible to characterize an interval only as a set of its center and range:

$$y = (c_y, r_y) \quad \text{with } c_y \in \mathbb{R}, r_y \in \mathbb{R}_0^+. \tag{2.6}$$

The use of these two notations is shown in the example below.

Example 2.2.2. *Considering an interval $y = [-2, 8]$, to obtain its representation with the centers and ranges notation in (2.5) and (2.6), we just need to do the following:*

- $c_y = \frac{-2+8}{2} = 3$,
- $r_y = 8 - (-2) = 10$,

thus, using the notation in (2.5), we obtain $y = [3 - \frac{10}{2}, 3 + \frac{10}{2}] = [-2, 8]$, and with the notation in (2.6) we have $y = (3, 10)$. □

Following the same reasoning, a histogram can also be characterized using centers and ranges. Taking a histogram x_j with n_j subintervals, it is possible to rewrite the bounds of the subintervals using their respective centers and ranges, such that:

$$x_j = \{c_{1j} - \frac{r_{1j}}{2}, c_{1j} + \frac{r_{1j}}{2}, p_{1j}; ...; c_{ij} - \frac{r_{ij}}{2}, c_{ij} + \frac{r_{ij}}{2}, p_{ij}; ...; c_{n_jj} - \frac{r_{n_jj}}{2}, c_{n_jj} + \frac{r_{n_jj}}{2}, p_{n_jj}\}. \tag{2.7}$$

Thus, 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 as the one represented in (2.7), the representation of its centers, ranges and probabilities is

$$x_j = (\boldsymbol{c}_j, \boldsymbol{r}_j, \boldsymbol{p}_j), \tag{2.8}$$

with

$$\boldsymbol{c}_j = \begin{bmatrix} c_{1j} \\ c_{2j} \\ \vdots \\ c_{ij} \\ \vdots \\ c_{n_j j} \end{bmatrix}, \boldsymbol{r}_j = \begin{bmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{ij} \\ \vdots \\ r_{n_j j} \end{bmatrix}, \boldsymbol{p}_j = \begin{bmatrix} p_{1j} \\ p_{2j} \\ \vdots \\ p_{ij} \\ \vdots \\ p_{n_j j} \end{bmatrix}.$$

The following example illustrates the use of the notation in (2.8) for a histogram.

Example 2.2.3. *Considering that we have the histogram x from Example 2.2.1.:*

$$x = \{[0, 3[, 0.1; [3, 4[, 0.3; [4, 8[, 0.4; [8, 15], 0.2\}.$$

To represent this histogram with the centers and ranges notation in (2.8), $x = (\boldsymbol{c}, \boldsymbol{r}, \boldsymbol{p})$, it is just necessary to calculate the vectors $\boldsymbol{c}$, $\boldsymbol{r}$, and $\boldsymbol{p}$:

$$\boldsymbol{c} = \begin{bmatrix} c_1 = \frac{0+3}{2} = \frac{3}{2} \\ c_2 = \frac{3+4}{2} = \frac{7}{2} \\ c_3 = \frac{4+8}{2} = 6 \\ c_4 = \frac{15+8}{2} = \frac{23}{2} \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} r_1 = 3 - 0 = 3 \\ r_2 = 4 - 3 = 1 \\ r_3 = 8 - 4 = 4 \\ r_4 = 15 - 8 = 7 \end{bmatrix}, \boldsymbol{p} = \begin{bmatrix} p_1 = 0.1 \\ p_2 = 0.3 \\ p_3 = 0.4 \\ p_4 = 0.2 \end{bmatrix}.$$

□

Another useful way to represent intervals and histograms is through a quantile function. This concept was introduced in [5] for interval variables and in [1] for histogram variables. Quantile functions were used as an important tool in some works, including [6], which proposes new linear regression models for interval and histogram data. This was a major source of information on how to deal with these functions for this thesis. The quantile function is related to the cumulative distribution function (cdf), F_X, of a random variable X, which is given by the expression

$$F_X(x) = P(X \leq x), \text{ with } x \in \mathbb{R}.$$

Therefore, a cdf gives the probability of a random variable X being less than or equal to a certain value x. The quantile function can be seen as the inverse of this function. 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 function Q_X for a random variable X with cdf $F_X(x)$ can be defined by

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

As it was stated previously, 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. Considering a uniformly distributed random variable Y defined over an interval $[a, b]$, its cdf, F_Y, is given by

$$F_Y(s) = \begin{cases} 0, & s < a \\ \frac{s-a}{b-a}, & s \in [a, b[\\ 1, & s \geq b \end{cases}. \tag{2.9}$$

Following the same reasoning, the quantile function, Q_Y, of a random variable Y with a uniform distribution defined over an interval $[a, b]$, according to [5], is

$$Q_Y(p) = a + (b - a)p,\ 0 \leq p \leq 1. \tag{2.10}$$

Using the center and ranges notation introduced previously, the equation (2.10) can also be rewritten as

$$Q_Y(p) = c_Y + \frac{r_Y}{2}(2p - 1),\ 0 \leq p \leq 1. \tag{2.11}$$

In Example 2.2.4, it is illustrated how an interval can be represented using the notations in (2.9), (2.10), and (2.11).

Example 2.2.4. *If we have the same interval as in Example 2.2.2., $y = [-2, 8]$, the associated cdf, F_y, can be obtained from (2.9):*

$$F_y(s) = \begin{cases} 0, & s < -2 \\ \frac{s+2}{10}, & s \in [-2, 8[\\ 1, & s \geq 8 \end{cases}.$$

The quantile function of y, Q_y, under the notation in (2.10), is given by

$$Q_y(p) = -2 + 10p,\ 0 \leq p \leq 1.$$

The quantile function Q_y can also be rewritten using the centers and ranges notation in (2.11). Knowing that, for this interval y, we have $c_y = 3$, and $r_y = 10$:

$$Q_y(p) = 3 + 10(2p - 1),\ 0 \leq p \leq 1.$$

□

The equations (2.9), (2.10), and (2.11) can also be applied to each of the subintervals of a histogram (since within a subinterval, micro-data follows a uniform distribution). Thus, it is also possible to create a representation for histograms using cdf's and quantile functions. The cdf F_X of the micro-data of a

histogram-valued variable X with n bins following the structure in (2.2), is given by

$$F_X(s) = \begin{cases} 0, & s \leq a_1 \\ \frac{s-a_1}{b_1-a_1}p_1, & a_1 \leq s < b_1 \\ \frac{s-a_2}{b_2-a_2}p_2, & a_2 \leq s < b_2 \\ \vdots & \\ \frac{s-a_i}{b_i-a_i}p_i, & a_i \leq s < b_i \\ \vdots & \\ 1, & s \geq b_n \end{cases}. \tag{2.12}$$

Knowing that the quantile function is the inverse of the cdf, we have that the quantile function, Q_X, of a histogram-valued variable X with n subintervals, in line with [1], is

$$Q_X(p) = \begin{cases} a_1 + \frac{p}{w_1}(b_1 - a_1), & 0 \leq p < w_1 \\ a_2 + \frac{p-w_1}{w_2-w_1}(b_2 - a_2), & w_1 \leq p < w_2 \\ \vdots & \\ a_i + \frac{p-w_{i-1}}{w_i-w_{i-1}}(b_i - a_i), & w_{i-1} \leq p < w_i \\ \vdots & \\ a_n + \frac{p-w_{n-1}}{1-w_{n-1}}(b_n - a_n), & w_{n-1} \leq p \leq 1 \end{cases}, \tag{2.13}$$

where the w_is are the cumulative probabilities of the first i subintervals, as represented in (2.2). Thus,

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

It is also useful to represent the quantile function of a histogram-valued variable X, Q_X, using the centers and ranges notation:

$$Q_X(p) = \begin{cases} c_1 + (\frac{2p}{w_1} - 1)\frac{r_1}{2}, & 0 \leq p < w_1 \\ c_2 + (\frac{2(p-w_1)}{w_2-w_1} - 1)\frac{r_2}{2}, & w_1 \leq p < w_2 \\ \vdots & \\ c_i + (\frac{2(p-w_{i-1})}{w_i-w_{i-1}} - 1)\frac{r_i}{2}, & w_{i-1} \leq p < w_i \\ \vdots & \\ c_n + (\frac{2(p-w_{n-1})}{1-w_{n-1}} - 1)\frac{r_n}{2}, & w_{n-1} \leq p \leq 1 \end{cases}. \tag{2.14}$$

The representation of a histogram according to the notations in (2.12), (2.13), and (2.14) is illustrated in the example that follows.

Example 2.2.5. *If we have a histogram x with three subintervals $(n = 3)$, which is represented in Figure 2.4, and has the following expression:*

$$x = \{[-3, 1[, 0.3; [1, 3[, 0.6; [3, 10], 0.1\}.$$

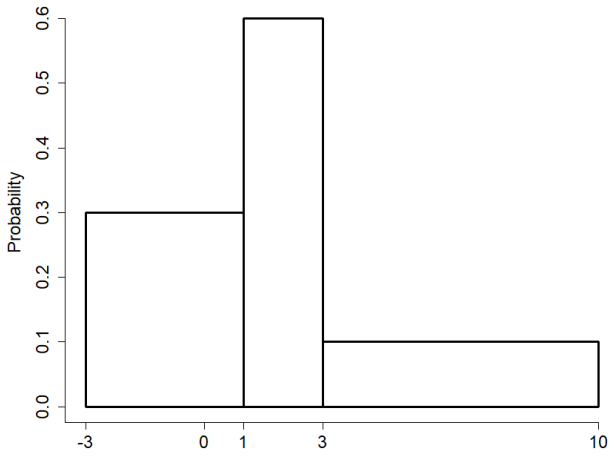


Figure 2.4: Histogram x associated with the Example 2.2.5.

The cdf of the micro-data associated with the histogram x, F_x, is defined as (see (2.12)):

$$F_x(s) = \begin{cases} 0, & s \leq -3 \\ \frac{s+3}{4} \times 0.3, & -3 \leq s < 1 \\ \frac{s-1}{2} \times 0.6, & 1 \leq s < 3 \\ \frac{s-3}{7} \times 0.1, & 3 \leq s < 10 \\ 1, & s \geq 10 \end{cases}.$$

Using the expression in (2.13), it is also possible to represent x by a quantile function. Firstly, it is necessary to calculate the cumulative probabilities w_i:

- $w_0 = 0$,
- $w_1 = p_1 = 0.3$,
- $w_2 = w_1 + p_2 = 0.9$,
- $w_3 = w_2 + p_3 = 1$.

This leads to the following quantile function for x, Q_x:

$$Q_x(p) = \begin{cases} -3 + \frac{p}{0.3} \times 4, & 0 \leq p < 0.3 \\ 1 + \frac{p-0.3}{0.6} \times 2, & 0.3 \leq p < 0.9 \\ 3 + \frac{p-0.9}{0.1} \times 7, & 0.9 \leq p \leq 1 \end{cases}.$$

To represent Q_x under the centers and ranges notation displayed in (2.14), it is necessary to obtain the centers and ranges for the three subintervals of this histogram x, $y_1 = [-3, 1[$, $y_2 = [1, 3[$, $y_3 = [3, 10]$:

- $c_1 = -1, r_1 = 4$,
- $c_2 = 2, r_2 = 2$,

- $c_3 = \frac{13}{2}, r_3 = 7$.

Therefore, the quantile function of x, Q_x, using centers and ranges, is given by

$$Q_x(p) = \begin{cases} -1 + (\frac{2p}{0.3} - 1) \times 2, & 0 \leq p < 0.3 \\ 2 + (\frac{2(p-0.3)}{0.6} - 1) \times 1, & 0.3 \leq p < 0.9 \\ \frac{13}{2} + (\frac{2(p-0.9)}{0.1} - 1) \times \frac{7}{2}, & 0.9 \leq p \leq 1 \end{cases}.$$

The graphic representation of $Q_x(p)$ is displayed in Figure 2.5, where it can be observed that this function is non-decreasing.

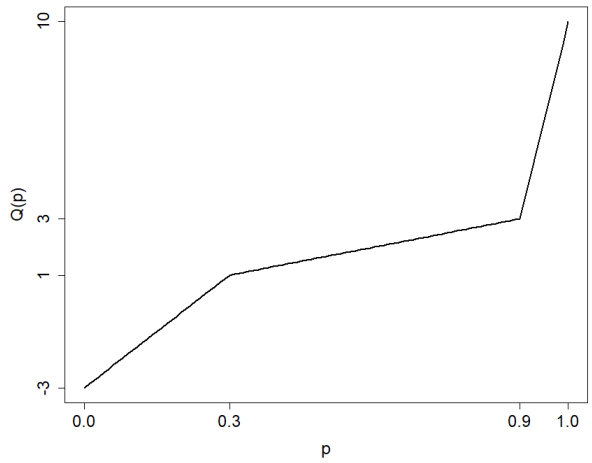


Figure 2.5: Graphic representation of $Q_x(p)$ from the Example 2.2.5.

□

A particular case occurs when one of the probabilities p_i associated with one of the subintervals of a histogram is equal to 0. When this happens, a discontinuity in the quantile function of the histogram is created, since we are unable to directly calculate $Q(w_{i-1})$. In this situation, it is only possible to calculate $\lim_{p \to w_{i-1}^+} Q(p)$ or $\lim_{p \to w_{i-1}^-} Q(p)$. In the following example, a case where this happens is analyzed.

Example 2.2.6. *Considering the following histogram x with three subintervals (displayed in Figure 2.6):*

$$x = \{[10, 20[, 0.8; [20, 30[, 0; [30, 45[, 0.2\}.$$

The second subinterval of this histogram has an associated probability $p_2 = 0$. Because of this, the value of x, for which the cdf of x is equal to p_1, can be any real number between the upper bound of the first interval (20) and the lower bound of the third interval (30). Considering that the quantile function $Q_x(p)$ is the inverse of the cdf $F_x(s)$, and that, for each value of p between 0 and 1, according to the definition of a function, the quantile function can only return one single value of x, $Q_x(p)$ is not continuous for this case at the point $p = w_1 = 0.8$. The quantile function for this example is, therefore:

$$Q_x(p) = \begin{cases} 10 + \frac{p}{0.8} \times 10, & 0 \leq p < 0.8 \\ 30 + \frac{p-0.8}{0.2} \times 15, & 0.8 < p \leq 1 \end{cases},$$

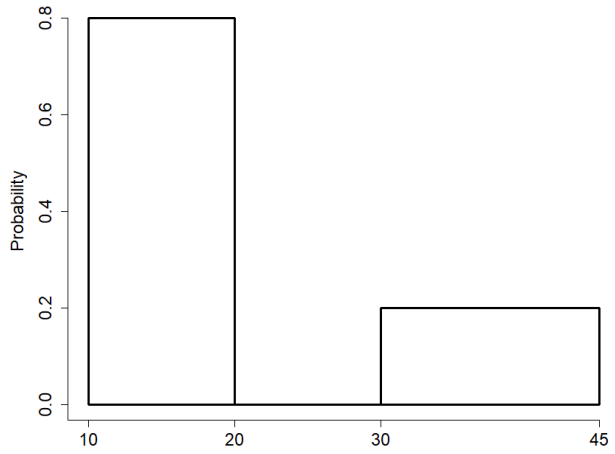


Figure 2.6: Histogram x associated with the Example 2.2.6.

and is not defined for $p = 0.8$.

The graphic representation of the quantile function $Q_x(p)$ *is displayed in Figure 2.7, where the discontinuity in the point* $p = 0.8$ *can be observed.*

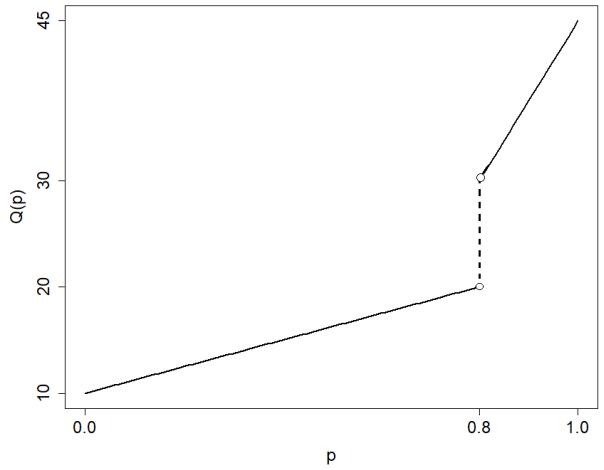


Figure 2.7: Graphic representation of $Q_x(p)$ from the Example 2.2.6.

□

From the previous examples it is possible to infer that the quantile function is always non-decreasing in its 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 ($a_{i+1} = b_i$). Another important property seen in the Example 2.2.6 is that this function can be non-continuous. The quantile functions introduced here are an important tool in the construction of an algebra for histograms, which is presented in the next chapter.

All the representations for intervals and histograms discussed in this chapter are also used in the following chapters. The choice of a specific representation depends on the convenience to illustrate the case under study.

Chapter 3

Interval and Histogram Algebra

In this chapter, the best known methods to do mathematical operations with both intervals and histograms are presented. For interval-valued variables, two algebras are discussed: Moore's Interval Algebra and the Extended Interval Algebra. For the case of the histogram-valued variables, an algebra based on arithmetic operations with quantile functions is also defined. This histogram algebra is used in Chapter 5 to propose a new estimation method for Symbolic Principal Components Analysis (SPCA), when it is applied to histogram-valued data.

3.1 Interval Algebras

Two of the most relevant interval algebras are Moore's Interval Algebra and the Extended Interval Algebra. Some of the properties of these algebras, as well as the advantages and disadvantages of their use, are presented in this sub-chapter.

3.1.1 Moore's Interval Algebra

The most commonly used algebra when dealing with interval variables is the one that was defined by Moore in [7]. This algebra will henceforth be called Moore's Interval Algebra.

Before proceeding with the definition of this algebra, it is important to clarify in which way the intervals are represented. To better understand how this algebra works, the operations are first described using the interval bounds notation in (2.1), and then represented as a set of its centers and ranges, as presented in (2.6). Thus, using the notation in (2.1), the i-th interval of a set of intervals $\mathcal{Y} = (y_1, .., y_i, .., y_n)$ is represented by

$$y_i = [a_i, b_i], \quad \text{with } a_i, b_i \in \mathbb{R}, a_i \leq b_i,$$

where a_i is the lower bound of the interval y_i, and b_i the upper bound.

Moore's Interval Algebra is defined over a set of intervals $\mathcal{Y}$ with the two following operations:

1. Addition: $y_1 + y_2 = [a_1, b_1] + [a_2, b_2] = [a_1 + a_2, b_1 + b_2]$.

2. Scalar multiplication: $\beta y_1 = \beta[a_1, b_1] = [min(\beta a_1, \beta b_1), max(\beta a_1, \beta b_1)], \forall \beta \in \mathbb{R}$.

By taking the particular case of the scalar multiplication, where $\beta = -1$, and subsequently performing an addition, it is also possible to define the difference between two intervals. The expression for the linear combination of two intervals can also be obtained by adding two intervals which were previously multiplied by two scalars. Therefore, the expressions for the difference and linear combination of intervals in Moore's Interval Algebra are:

- Difference: $y_1 - y_2 = [a_1, b_1] - [a_2, b_2] = [a_1, b_1] + [min(-a_2, -b_2), max(-a_2, -b_2)] =$
  $= [a_1 + min(-a_2, -b_2), a_2 + max(-a_2, -b_2)]$.

- Linear combination: $\beta_1 y_1 + \beta_2 y_2 = \beta_1[a_1, b_1] + \beta_2[a_2, b_2] = [min(\beta_1 a_1, \beta_1 b_1), max(\beta_1 a_1, \beta_1 b_1)]$
  $+ [min(\beta_2 a_2, \beta_2 b_2), max(\beta_2 a_2, \beta_2 b_2)] = [min(\beta_1 a_1, \beta_1 b_1) + min(\beta_2 a_2, \beta_2 b_2),$
  $, max(\beta_1 a_1, \beta_1 b_1) + max(\beta_2 a_2, \beta_2 b_2)], \forall \beta_1, \beta_2 \in \mathbb{R}$.

It is worth noting that all these previous operations in Moore's Interval Algebra guarantee that the lower bounds of the resulting intervals $y_i = [a_i, b_i]$ always remain smaller than their respective upper bounds $(a_i < b_i)$.

When using this representation with the interval bounds, the expression for the linear combination between intervals is not very simple. For this reason, we prefer to use the centers and ranges notation to simplify these results. This approach allows us to reach a simplified general formula for the linear combinations with Moore's Interval Algebra. Thus, using the representation (2.6) for intervals, each element y_i belonging to the set $\mathcal{Y}$ of interval variables has the following representation:

$$y_i = (c_i, r_i), \quad \text{with } c_i \in \mathbb{R}, r_i \in \mathbb{R}_0^+,$$

where $c_i = \frac{a_i + b_i}{2}$ and $r_i = b_i - a_i$.

To perform intermediate calculations, it is useful to first characterize the bounds of the intervals y_i with the center and range notation (2.5):

$$y_i = [c_i - \tfrac{r_i}{2}, c_i + \tfrac{r_i}{2}].$$

Under these conditions, the four previously described operations of Moore's Interval Algebra, according to the notation with the bounds of the subintervals in (2.5) (in square brackets) and the notation with the set of centers and ranges in (2.6) (in parentheses), take the following form:

1. Addition: $y_1 + y_2 = [c_1 + c_2 - \frac{r_1 + r_2}{2}, c_1 + c_2 + \frac{r_1 + r_2}{2}] = (c_1 + c_2, r_1 + r_2)$.

2. Scalar multiplication: $\beta y_1 = [\beta c_1 - |\beta| \frac{r_1}{2}, \beta c_1 + |\beta| \frac{r_1}{2}] = (\beta c_1, |\beta| r_1), \forall \beta \in \mathbb{R}$.

3. Difference: $y_1 - y_2 = [c_1 - c_2 - \frac{r_1 + r_2}{2}, c_1 - c_2 + \frac{r_1 + r_2}{2}] = (c_1 - c_2, r_1 + r_2)$.

4. Linear combination: $\beta_1 y_1 + \beta_2 y_2 = [\beta_1 c_1 + \beta_2 c_2 - \frac{|\beta_1| r_1 + |\beta_2| r_2}{2}, \beta_1 c_1 + \beta_2 c_2 + \frac{|\beta_1| r_1 + |\beta_2| r_2}{2}] =$
   $= (\beta_1 c_1 + \beta_2 c_2, |\beta_1| r_1 + |\beta_2| r_2), \forall \beta_1, \beta_2 \in \mathbb{R}$.

The results for the operations of Moore's Interval Algebra are simpler and easier to interpret. The information regarding the resulting centers and ranges of the intervals for these operations is summarized in Table 3.1. In this table, the operation of addition of an interval with a constant is introduced ($y_1 + \beta$). To do this operation, it is considered that the constant β corresponds to an interval where the lower and upper bounds are the same. That is, for a value β, we make operations using the interval [β, β] with the bounds notation. This corresponds to an interval with a center equal to β, and a range equal to 0. Therefore, its representation under the centers and ranges notation is $(\beta, 0)$.

Table 3.1: 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	$\lvert\beta\rvert r_1$
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\lvert\beta_1\rvert r_1 + \lvert\beta_2\rvert r_2$

As observed before, since the range is always positive in all the operations of Moore's Interval Algebra, it is guaranteed that the resulting intervals continue to follow the condition of the upper bound being greater than or equal to the lower bound. However, this causes a disadvantage: the ranges of the resulting intervals keep increasing with each consecutive operation. This could lead to very large intervals, which may lose their significance. It is also interesting to note that, for this reason, this algebra cannot be considered a vector space, since the additive inverse axiom fails to hold. This axiom requires that, when considering a set of intervals $\mathcal{Y} = (y_1, .., y_i, .., y_n)$, for every $y_i \in \mathcal{Y}$, there exists an element $y_i \in \mathcal{Y}$, such that $y_i + (-y_i) = 0$. Since the ranges of the resulting intervals can never decrease and, thus, go back to 0, this condition is not satisfied in Moore's Interval Algebra. Instead, the operation $y_i - y_i$ results in an interval with center 0 and with a range that is the double of the original range of y_i.

The operations of this algebra are illustrated in Example 3.1.1.

Example 3.1.1. *Considering the intervals $y_1 = [2, 6]$ and $y_2 = [-2, 3]$, if we want to do the operations in Moore's Interval Algebra, we can first convert them into a set of the centers and ranges (see 2.6), do the operations according to the rules in Table 3.1 and then convert them back to the bounds notation in (2.1).*

Calculating the centers and ranges of y_1 and y_2, we obtain: $c_1 = 4$, $r_1 = 4$, $c_2 = \frac{1}{2}$ $r_2 = 5$. Thus, under the centers and ranges notation, $y_1 = (4, 4)$ and $y_2 = (\frac{1}{2}, 5)$. Knowing this, it is possible to easily perform the following list of operations:

- $y_1 + 3 = (4, 4) + (3, 0) = (7, 4) = [5, 9]$;
- $y_1 + y_2 = (4, 4) + (\frac{1}{2}, 5) = (\frac{9}{2}, 9) = [0, 9]$;
- $2 \times y_1 = 2 \times (4, 4) = (8, 8) = [4, 12]$;

- $-3 \times y_1 = -3 \times (4,4) = (-12,12) = [-18,-6]$;
- $y_1 - y_1 = (4,4) - (4,4) = (0,8) = [-4,4]$;
- $y_1 - y_2 = (4,4) - (\frac{1}{2},5) = (\frac{7}{2},9) = [-1,8]$;
- $-2 \times y_1 + 3 \times y_2 = -2 \times (4,4) + 3 \times (\frac{1}{2},5) = (-8,8) + (\frac{3}{2},15) = (-\frac{13}{2},23) = [-18,5]$.

□

It is also possible to find a generalized formula for the linear combinations with Moore's Interval Algebra. This is achieved by aggregating in vectors the sets of the centers and ranges of the corresponding intervals, and also the scalars β. Considering that we are doing linear combinations in a set of n intervals y_i, the following $\boldsymbol{y}$, $\boldsymbol{c}$, $\boldsymbol{r}$ and $\boldsymbol{\beta}$ vectors are accordingly created:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_i \\ \vdots \\ c_n \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_n \end{bmatrix}.$$

Using these vectors, the general expression for the linear combination of intervals with Moore's Interval Algebra is given by

$$\boldsymbol{\beta}^t \boldsymbol{y} = [\boldsymbol{\beta}^t \boldsymbol{c} - |\boldsymbol{\beta}|^t \frac{\boldsymbol{r}}{2},\ \boldsymbol{\beta}^t \boldsymbol{c} + |\boldsymbol{\beta}|^t \frac{\boldsymbol{r}}{2}], \tag{3.1}$$

where $\boldsymbol{\beta}^t$ is the transpose of $\boldsymbol{\beta}$ and $|\boldsymbol{\beta}|$ corresponds to the absolute value of $\boldsymbol{\beta}$.

Expression (3.1) can also be represented by considering the resulting interval to be a set of its centers and ranges:

$$\boldsymbol{\beta}^t \boldsymbol{y} = (\boldsymbol{\beta}^t \boldsymbol{c},\ |\boldsymbol{\beta}|^t \boldsymbol{r}). \tag{3.2}$$

In the example below, it is shown how the general expression in (3.1) can be used to compute a linear combination of intervals.

Example 3.1.2. *Considering that we have the set of four intervals:*

$$y_1 = [-5,-2],\ y_2 = [0,3],\ y_3 = [-2,4],\ y_4 = [1,9],$$

and we want to calculate the following linear combination of these intervals,

$$\widetilde{y} = 2y_1 - 3y_2 + 10y_3 - 4y_4,$$

the first step is to obtain the previously mentioned vectors with the aggregated intervals, centers, ranges and constants, $\boldsymbol{y}$, $\boldsymbol{c}$, $\boldsymbol{r}$, *and* $\boldsymbol{\beta}$, *respectively:*

$$\boldsymbol{y} = \begin{bmatrix} [-5,-2] \\ [0,3] \\ [-2,4] \\ [1,9] \end{bmatrix}, \boldsymbol{c} = \begin{bmatrix} -\frac{7}{2} \\ \frac{3}{2} \\ 1 \\ 5 \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} 3 \\ 3 \\ 6 \\ 8 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}.$$

Now, using equation (3.1), we can calculate the interval $\widetilde{y}$:

$$\widetilde{y} = \boldsymbol{\beta}^t \boldsymbol{y} = \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} [-5,-2] \\ [0,3] \\ [-2,4] \\ [1,9] \end{bmatrix} = \left[\begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} -\frac{7}{2} \\ \frac{3}{2} \\ 1 \\ 5 \end{bmatrix} - \frac{1}{2} \left| \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \right| \begin{bmatrix} 3 \\ 3 \\ 6 \\ 8 \end{bmatrix}, \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} -\frac{7}{2} \\ \frac{3}{2} \\ 1 \\ 5 \end{bmatrix} + \frac{1}{2} \left| \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \right| \begin{bmatrix} 3 \\ 3 \\ 6 \\ 8 \end{bmatrix} \right] \Leftrightarrow$$

$$\Leftrightarrow \widetilde{y} = [-75, 32].$$

□

3.1.2 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 [8] and was used in some works related with time series [9] and also with fuzzy numbers [10]. Henceforth, it will be designated by Extended Interval Algebra.

The addition, scalar multiplication, difference and linear combination operations for intervals are defined in the following way for the Extended Interval Algebra, when the notation (2.1) with the bounds of the intervals is used:

1. Addition: $y_1 + y_2 = [a_1, b_1] + [a_2, b_2] = [a_1 + a_2, b_1 + b_2]$.
2. Scalar multiplication: $\beta y_1 = \beta[a_1, b_1] = [\beta a_1, \beta b_1], \forall \beta \in \mathbb{R}$.
3. Difference: $y_1 - y_2 = [a_1, b_1] - [a_2, b_2] = [a_1 - b_1, a_2 - b_2]$.
4. Linear combination: $\beta_1 y_1 + \beta_2 y_2 = \beta_1[a_1, b_1] + \beta_2[a_2, b_2] = [\beta_1 a_1 + \beta_2 a_2, \beta_1 a_2 + \beta_2 b_2], \forall \beta_1, \beta_2 \in \mathbb{R}$.

Consequently, while the addition operation is the same as in Moore's, the three other operations are different and no longer guarantee that the condition of the value of the lower bound of an interval being smaller than its upper bound is met. This originates a new type of interval, usually denominated by an Extended Interval, which is defined according to the following expression:

$$y = [a, b] \quad \text{with } a \in \mathbb{R}, b \in \mathbb{R}. \tag{3.3}$$

In the previous expression (3.3), the condition $a < b$ is no longer included as in (2.1). Hence, for example, the operation $-2 \times [1, 2]$, for this algebra, would originate the extended interval $[-2, -4]$, where $a = -2 > b = -4$.

The operations of the Extended Interval Algebra, under the centers and ranges notation, are:

1. Addition: $y_1 + y_2 = [c_1 + c_2 - \frac{r_1 + r_2}{2}, c_1 + c_2 + \frac{r_1 + r_2}{2}] = (c_1 + c_2, r_1 + r_2)$.
2. Scalar multiplication: $\beta y_1 = [\beta c_1 - \beta \frac{r_1}{2}, \beta c_1 + \beta \frac{r_1}{2}] = (\beta c_1, \beta r_1), \forall \beta \in \mathbb{R}$.
3. Difference: $y_1 - y_2 = [c_1 - c_2 - \frac{r_1 - r_2}{2}, c_1 - c_2 + \frac{r_1 - r_2}{2}] = (c_1 - c_2, r_1 - r_2)$.

4. Linear combination: $\beta_1 y_1 + \beta_2 y_2 = [\beta_1 c_1 + \beta_2 c_2 - \frac{\beta_1 r_1 + \beta_2 r_2}{2}, \beta_1 c_1 + \beta_2 c_2 + \frac{\beta_1 r_1 + \beta_2 r_2}{2}] =$
$= (\beta_1 c_1 + \beta_2 c_2, \beta_1 r_1 + \beta_2 r_2), \forall \beta_1, \beta_2 \in \mathbb{R}$.

Using this information, it is possible to create a table with the resulting centers and ranges for each of these operations for the Extended Interval Algebra. These results are displayed in Table 3.2. 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. As a consequence, 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 3.2: 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
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\beta_1 r_1 + \beta_2 r_2$

The operations of this algebra are illustrated in the example that follows.

Example 3.1.3. *Considering the intervals $y_1 = [2, 6]$ and $y_2 = [-2, 3]$ and operations from Example 3.1.2, but now doing them according to the Extended Interval Algebra, the results obtained are:*

- $y_1 + 3 = (4, 4) + (3, 0) = (7, 4) = [5, 9]$;
- $y_1 + y_2 = (4, 4) + (\frac{1}{2}, 5) = (\frac{9}{2}, 9) = [0, 9]$;
- $2 \times y_1 = 2 \times (4, 4) = (8, 8) = [4, 12]$;
- $-3 \times y_1 = -3 \times (4, 4) = (-12, -12) = [-6, -18]$;
- $y_1 - y_1 = (4, 4) - (4, 4) = (0, 0) = [0, 0] = 0$;
- $y_1 - y_2 = (4, 4) - (\frac{1}{2}, 5) = (\frac{7}{2}, -1) = [3, 4]$;
- $-2 \times y_1 + 3 \times y_2 = -2 \times (4, 4) + 3 \times (\frac{1}{2}, 5) = (-8, -8) + (\frac{3}{2}, 15) = (-\frac{13}{2}, 7) = [-10, -3]$.

Comparing with the results obtained with Moore's algebra, it can be concluded that, while there are similar results for the first three operations where only positive β's are involved, significant differences are observed when at least a negative weight β is included in the operation. For the particular case of the difference between two equal intervals ($y_1 - y_1$), the result is always equal to 0 in this algebra. □

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:

$$\boldsymbol{\beta}^t \boldsymbol{y} = [\boldsymbol{\beta}^t \boldsymbol{c} - \boldsymbol{\beta}^t \frac{\boldsymbol{r}}{2},\ \boldsymbol{\beta}^t \boldsymbol{c} + \boldsymbol{\beta}^t \frac{\boldsymbol{r}}{2}]. \tag{3.4}$$

And considering the notation where the resulting interval is a a set of its centers and ranges:

$$\boldsymbol{\beta}^t\boldsymbol{y} = (\boldsymbol{\beta}^t\boldsymbol{c},\ \boldsymbol{\beta}^t\boldsymbol{r}). \tag{3.5}$$

An application of the expression (3.4) to compute the linear combination of intervals is shown in Example 3.1.4.

Example 3.1.4. *Considering the same conditions presented in Example 3.1.2., but now calculating the linear combination according to the Extended Interval Algebra, the result is:*

$$y = \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix} \begin{bmatrix} [-5,-2] \\ [0,3] \\ [-2,4] \\ [1,9] \end{bmatrix} = \left[\begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} -\frac{7}{2} \\ \frac{3}{2} \\ 1 \\ 5 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} 3 \\ 3 \\ 6 \\ 8 \end{bmatrix}, \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} -\frac{7}{2} \\ \frac{3}{2} \\ 1 \\ 5 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 2 \\ -3 \\ 10 \\ -4 \end{bmatrix}^t \begin{bmatrix} 3 \\ 3 \\ 6 \\ 8 \end{bmatrix} \right] \Leftrightarrow$$

$$\Leftrightarrow y = [-34, -9].$$

□

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. Moreover, that representation makes no sense for the symbolic data sets that we deal with, whose interval-valued variables always respect the condition $a \leq b$. That is why it was decided for this work that, when this case occurs for this algebra, the upper and lower bounds are switched and, consequently, a regular interval is created. For instance, if an extended interval $[9, 3]$ is generated with this algebra, the 9 and 3 are switched and the final result is the interval $[3, 9]$.

With this alteration, the new results of the centers and ranges for the different operations are presented in Table 3.3. The only difference when comparing them with Table 3.2 is that, now, the values of the ranges always correspond to an absolute value and, consequently, there can be no negative ranges. Since $|\beta_1 r_1 + \beta_2 r_2| \leq |\beta_1| r_1 + |\beta_2| r_2$, the intervals resulting from a linear combination using the Extended Interval Algebra are always included in the intervals obtained from the same linear combination when Moore's Interval Algebra is used.

Table 3.3: 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	$\lvert \beta r_1 \rvert$
$\beta_1 y_1 + \beta_2 y_2$	$\beta_1 c_1 + \beta_2 c_2$	$\lvert \beta_1 r_1 + \beta_2 r_2 \rvert$

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, 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 have no usefulness 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, even though it is not a linear space, given that in this algebra $y_1 - y_1 \neq 0$.

3.2 Histogram operations with quantile functions

Developing an algebra for histogram-valued variables is more complex than for the interval case. This is mainly because, with histograms, we have to deal with several subintervals, each one with an associated probability p_i. These probabilities p_i make it difficult to find a way to relate the subintervals from different histograms, in order to be able to make arithmetic operations with them. Furthermore, the fact that each histogram can have a different number of subintervals also further complicates matters.

In [11], Colombo and Jaarma developed a method to do operations with histograms, but the results lead to histograms with subintervals which could not be ordered and would intersect each other. Methods to rearrange these subintervals were also developed by the same authors, but the procedure was too complex and the results were not ideal because the distribution of the subintervals was not considered in the resulting representation of the histograms. For this reason, it was necessary to develop a way to perform arithmetic operations with better results. This was obtained by Irpino and Verde in [1], using the quantile functions previously mentioned in Chapter 2, to perform arithmetic operations with histograms, which is analyzed in this sub-chapter.

The definition of quantile function for histograms in (2.13) is the basis for these operations. However, these quantiles cannot be used directly in the operations and a transformation of the histograms has to be made beforehand. This step consists in transforming the original histograms into histograms where the number of subintervals is the same and where the branches of the quantile functions are defined over the same intervals of cumulative probabilities.

3.2.1 Harmonization

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 first 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. In this way, it is easier to perform the operations with quantile functions. This procedure was introduced in [1] and, in our work, it 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 make 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. Thus, if $\mathcal{W}_j$ is considered as the set of cumulative probabilities for the histogram j, w_{ij} the i-th cumulative probability of the histogram j, and n_j the number of original subintervals for the histogram j, from a total of k histograms, we would have the following set $\mathcal{W}$:

$$\mathcal{W} = \{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_j, ..., \mathcal{W}_k\},$$

with:

$$\begin{aligned}
\mathcal{W}_1 &= \{w_{01}, w_{11}, w_{21}, ..., w_{i1}, ..., w_{n_1 1}\},\\
\mathcal{W}_2 &= \{w_{02}, w_{12}, w_{22}, ..., w_{i2}, ..., w_{n_2 2}\},\\
&\vdots\\
\mathcal{W}_j &= \{w_{0j}, w_{1j}, w_{2j}, ..., w_{ij}, ..., w_{n_j j}\},\\
&\vdots\\
\mathcal{W}_k &= \{w_{0k}, w_{1k}, w_{2k}, ..., w_{ik}, ..., w_{n_k k}\}.
\end{aligned}$$

The next step of this procedure is to merge all these $\mathcal{W}_j$ into a single set of cumulative probabilities with n elements. In our work, this set is denoted by $\mathcal{W}^\star$:

$$\mathcal{W}^\star = \{w_0^\star, w_1^\star, w_2^\star, ..., w_i^\star, ..., w_n^\star\}.$$

This new set $\mathcal{W}^\star$ gathers all the w_{ij} from the previously defined sets $\mathcal{W}_j$, so that the resulting $w_i^\star$s are sorted in ascending order and none of its elements are repeated. The size of this set $\mathcal{W}^\star$ is n, which corresponds to the number of non-repeated different elements from all the sets $\mathcal{W}_j$. 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. $max\{n_1, n_2, .., n_k\} \leq n \leq \sum_{j=1}^{k} n_j - 1$.

In the two following examples, it is shown how to compute the set of cumulative probabilities $\mathcal{W}^\star$ for two different cases.

Example 3.2.1. *Considering two histograms x_1 and x_2, which can be represented with the notation in (2.2) as*

$$\begin{aligned}
x_1 &= \{[-5, -2[, 0.5; [-2, 3[, 0.3; [3, 4], 0.2\},\\
x_2 &= \{[3, 10[, 0.4; [10, 12], 0.6\}.
\end{aligned}$$

x_1 and x_2 can also be represented with a quantile function, according to (2.13), as follows:

$$Q_{x_1}(p)= \begin{cases} -5 + \frac{p}{0.5} \times 3, & 0 \leq p < 0.5 \\ -2 + \frac{p-0.5}{0.3} \times 5, & 0.5 \leq p < 0.8 \\ 3 + \frac{p-0.8}{0.2}, & 0.8 \leq p \leq 1 \end{cases},$$

$$Q_{x_2}(p)= \begin{cases} 3 + \frac{p}{0.4} \times 7, & 0 \leq p < 0.4 \\ 10 + \frac{p-0.4}{0.6} \times 2, & 0.4 \leq p \leq 1 \end{cases}.$$

The histogram x_1 has three subintervals, while x_2 has two. The sets $\mathcal{W}_1$ and $\mathcal{W}_2$ for the cumulative probabilities of the histograms x_1 and x_2 are, respectively:

$$\mathcal{W}_1 = \{0, 0.5, 0.8, 1\},$$
$$\mathcal{W}_2 = \{0, 0.4, 1\}.$$

The only elements that are repeated in both $\mathcal{W}_1$ and $\mathcal{W}_2$ are w_{01} and w_{02}, which are both 0, along with w_{31} and w_{22}, both equal to 1. This is to be expected, since the first and last elements of the sets of cumulative probabilities must always be 0 and 1, respectively. All the other elements of both sets are different from each other. Thus, when creating the new set of cumulative probabilities, $\mathcal{W}^\star$, all the elements from both sets are taken, but only one of the zeros and one of the ones are included. This generates the following set $\mathcal{W}^\star$ with five elements:

$$\mathcal{W}^\star = \{0, 0.4, 0.5, 0.8, 1\}.$$

□

Example 3.2.2. *Considering we have three histograms with the following set of cumulative probabibilities:*

$$\mathcal{W}_1 = \{0, 0.1, 0.3, 0.7, 1\},$$
$$\mathcal{W}_2 = \{0, 0.3, 0.7, 1\},$$
$$\mathcal{W}_3 = \{0, 0.1, 1\},$$

then, all the elements from $\mathcal{W}_2$ and $\mathcal{W}_3$ would be repeated in $\mathcal{W}_1$. Hence, the set $\mathcal{W}^\star$ would be equal to the set $\mathcal{W}_1$:

$$\mathcal{W}^\star = \mathcal{W}_1 = \{0, 0.1, 0.3, 0.7, 1\}.$$

□

Now that a new set of harmonized cumulative probabilities $\mathcal{W}^\star$ has been defined, the next step in the harmonization process is to build new quantile functions for the histograms using the newly defined $w_i^\star$s as the new bounds of the branches of the function. Therefore, the subintervals that define the branches of the harmonized quantile function now have the form $[w_{i-1}^\star, w_i^\star[$ with $i \in \{1, .., n\}$. By doing this, the subintervals of the original histograms are being split according to the new cumulative probabilities $w_i^\star$. To compute the bounds of the new subintervals for each histogram x_j, it is necessary to calculate 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 quantile function, $Q_{x_j^\star}$, using the new set of cumulative probabilities $\mathcal{W}^\star$ for a histogram x_j, is given by

$$Q_{x_j^\star}(p) = \begin{cases} Q_{x_j}(w_0^\star) + \frac{p}{w_1^\star}(Q_{x_j}(w_1^\star) - Q_{x_j}(w_0^\star)), & 0 \leq p < w_1^\star \\ Q_{x_j}(w_1^\star) + \frac{p-w_1^\star}{w_2^\star - w_1^\star}(Q_{x_j}(w_2^\star) - Q_{x_j}(w_1^\star)), & w_1^\star \leq p < w_2^\star \\ \vdots & \\ Q_{x_j}(w_{i-1}^\star) + \frac{p-w_{i-1}^\star}{w_i^\star - w_{i-1}^\star}(Q_{x_j}(w_i^\star) - Q_{x_j}(w_{i-1}^\star)), & w_{i-1}^\star \leq p < w_i^\star \\ \vdots & \\ Q_{x_j}(w_{n-1}^\star) + \frac{p-w_{n-1}^\star}{1-w_{n-1}^\star}(Q_{x_j}(w_n^\star) - Q_{x_j}(w_{n-1}^\star)), & w_{n-1}^\star \leq p \leq 1 \end{cases}. \tag{3.6}$$

Now the process of harmonization of the histograms is completed and the resulting harmonized histograms have their respective quantile functions defined over the same subintervals of cumulative probabilities $w_i^\star$. This allows us to perform mathematical operations using these harmonized quantile functions easily.

It is also possible to represent a histogram $x_j^\star$ that went through the process of harmonization with the structure in (2.2) as

$$x_j^\star = \{[Q_{x_j}(w_0^\star), Q_{x_j}(w_1^\star)[, p_1^\star; ...; [Q_{x_j}(w_{i-1}^\star), Q_{x_j}(w_i^\star)[, p_i^\star; ...; [Q_{x_j}(w_{n-1}^\star), Q_{x_j}(w_n^\star)], p_n^\star\}. \tag{3.7}$$

Since all the histograms that went through this process are defined over the same set of cumulative probabilities $\mathcal{W}^\star$, and, consequently, the same number of n subintervals, it also means that, for each histogram j, the set of probabilities $p_{ij}^\star$ is the same. Therefore, the index j from these probabilities can be removed and the harmonized probabilities can be represented as $p_i^\star$, where i represents the subinterval it is associated with. These newly formed $p_i^\star$ can easily be computed from the harmonized cumulative probabilities $w_i^\star$ through the expression:

$$p_i^\star = w_i^\star - w_{i-1}^\star \text{ with } i \in \{1, .., n\}.$$

All the steps of the harmonization procedure are illustrated in Example 3.2.3.

Example 3.2.3. *Considering that we have the histograms x_1 and x_2 from Example 3.2.1 with the harmonized cumulative probability set $\mathcal{W}^\star = \{0, 0.4, 0.5, 0.8, 1\}$, the harmonized quantile functions, $Q_{x_1^\star}$ and $Q_{x_2^\star}$, for x_1 and x_2, respectively, are now calculated.*

As, for x_1, the only element of $\mathcal{W}^\star$ which did not belong to its original set $\mathcal{W}_1$ is 0.4, it is just necessary to calculate $Q_{x_1}(0.4)$ and then rearrange the quantile function accordingly.

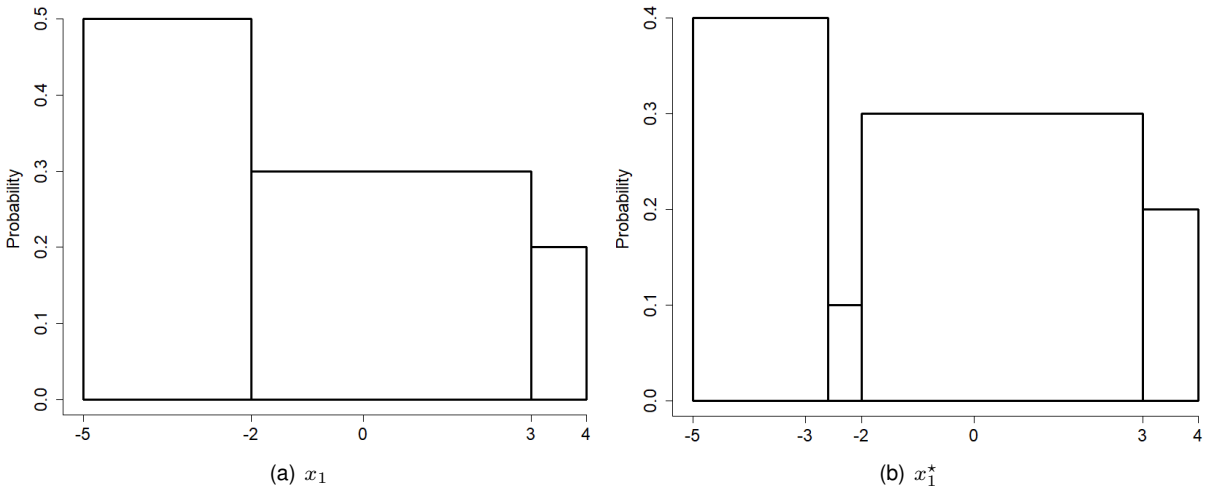
We have that $Q_{x_1}(0.4) = -5 + \frac{0.4}{0.5} \times 3 = -\frac{13}{5}$. Thus, the first original subinterval $[-5, -2[$ of x_1 was split into two new subintervals: $[-5, -\frac{13}{5}[$ and $[-\frac{13}{5}, -2]$ with an associated probability of $p_1^\star = w_1^\star - w_0^\star = 0.4$, and $p_2^\star = w_2^\star - w_1^\star = 0.5 - 0.4 = 0.1$, respectively. The other two remaining subintervals suffer no alterations. Now, it is just necessary to rearrange the quantile function to consider these new subintervals, thus creating the harmonized quantile function $Q_{x_1^\star}$:

$$Q_{x_1^\star}(p)= \begin{cases} -5+\frac{p}{0.4}\times\frac{12}{5}, & 0\leq p<0.4 \\ -\frac{13}{5}+\frac{p-0.4}{0.1}\times\frac{3}{5}, & 0.4\leq p<0.5 \\ -2+\frac{p-0.5}{0.3}\times 5, & 0.5\leq p<0.8 \\ 3+\frac{p-0.8}{0.2}, & 0.8\leq p\leq 1 \end{cases}.$$

The newly created harmonized histogram $x_1^\star$ can also be represented with the notation in (3.3):

$$x_1^\star=\{[-5,-\tfrac{13}{5}[,0.4;[-\tfrac{13}{5},-2[,0.1;[-2,3[,0.3;[3,4],0.2\}.$$

Figure 3.1 displays a comparison between the original histogram x_1 and its counterpart $x_1^\star$, originated from the previously mentioned harmonization process. As mentioned previously, the histogram remained practically the same, except for the first subinterval, which was divided into two.



(a) x_1 (b) $x_1^\star$

Figure 3.1: Original histogram x_1 and its harmonized version $x_1^\star$ from Example 3.2.3.

For the histogram x_2, two elements are a part of the harmonized set $\mathcal{W}^\star$ and are not included in the original set $\mathcal{W}_2$: 0.5 and 0.8. Thus, it is necessary to calculate $Q_{x_2}(0.5)$ and $Q_{x_2}(0.8)$ to proceed.

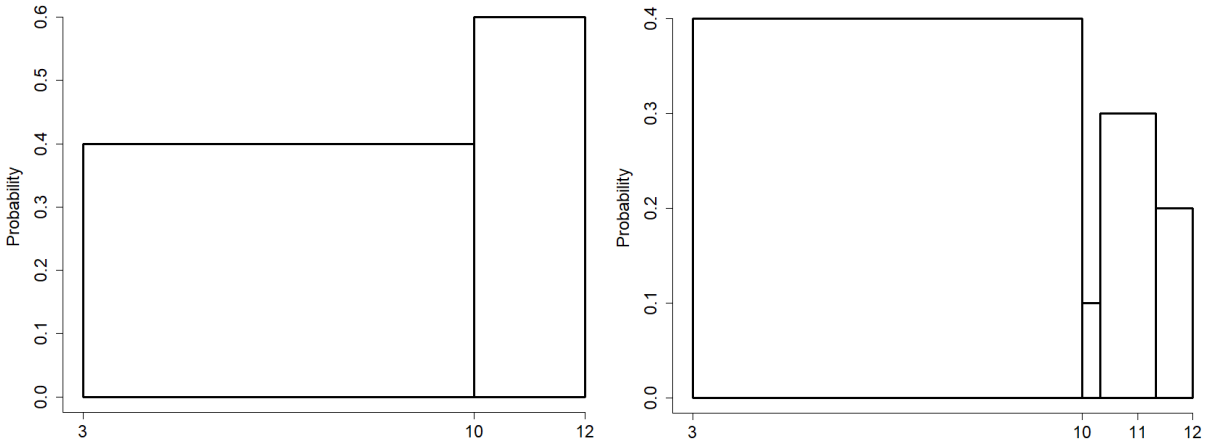
We have that $Q_{x_2}(0.5)=10+\frac{0.5-0.4}{0.6}\times 2=\frac{31}{3}$ and $Q_{x_2}(0.8)=10+\frac{0.8-0.4}{0.6}\times 2=\frac{34}{3}$. Therefore, the first subinterval of x_2 $[3,10[$ remains unchanged, while the second original subinterval $[10,12]$ is now split into three subintervals: $[10,\frac{31}{3}[$, $[\frac{31}{3},\frac{34}{3}[$, and $[\frac{34}{3},12]$ with associated probabilities of $p_2^\star=w_2^\star-w_1^\star=0.5-0.4=0.1$, $p_3^\star=w_3^\star-w_2^\star=0.8-0.5=0.3$, $p_4^\star=w_4^\star-w_3^\star=1-0.8=0.2$, respectively. This gives us the following harmonized quantile function for x_2, $Q_{x_2^\star}$:

$$Q_{x_2^\star}(p)= \begin{cases} 3+\frac{p}{0.4}\times 7, & 0\leq p<0.4 \\ 10+\frac{p-0.4}{0.1}\times\frac{1}{3}, & 0.4\leq p<0.5 \\ \frac{31}{3}+\frac{p-0.5}{0.3}, & 0.5\leq p<0.8 \\ \frac{34}{3}+\frac{p-0.8}{0.2}\times\frac{2}{3}, & 0.8\leq p\leq 1 \end{cases}.$$

Also representing this harmonized x_2 in the notation from (3.3), we have

$$x_2^\star=\{[3,10[,0.4;[10,\frac{31}{3}[,0.1;[\frac{31}{3},\frac{34}{3}[,0.3;[\frac{34}{3},12],0.2\}.$$

In Figure 3.2, both the original histogram x_2 with only two subintervals and the resulting harmonized histogram $x_2^\star$ with five subintervals are represented. It can be observed that the three subintervals of $x_2^\star$ which resulted from the division of the second subinterval of x_2 are very small.



(a) x_2 (b) $x_2^\star$

Figure 3.2: Original histogram x_2 and its harmonized version $x_2^\star$ from Example 3.2.3.

□

As it can be perceived from the previous example, 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. There is a solution to avoid this issue, but it can only be applied in the case where the histograms used have their micro-data available. The solution consists of manually building the macro-data, thus guaranteeing that all the subintervals of the histograms are defined within the same range of cumulative of probabilities. By doing so, the harmonization process would be useless. However, this option is not valid for many cases, since the micro-data is not always available. Furthermore, if we are not dealing with extreme cases, the harmonization process is a fairly good option to perform mathematical operations with histogram variables.

3.2.2 Arithmetic Operations

After having performed the harmonization procedure described previously, we have k histograms with an equal number of subintervals n, which are defined over the same set of harmonized cumulative probabilities $\mathcal{W}^\star$. As a result, it is easy to perform arithmetic operations with these harmonized histograms using their harmonized quantile functions, according to the method proposed in [1].

Under these conditions, the sum of two histograms is done by simply adding their harmonized quantile functions. Therefore, if we have two histograms x_1 and x_2, which were subsequently harmonized into the histograms $x_1^\star$ and $x_2^\star$ with the following expressions:

$$x_1^\star = \{[a_{11}^\star, b_{11}^\star[, p_1^\star; ...; [a_{i1}^\star, b_{i1}^\star[, p_i^\star; ...; [a_{n1}^\star, b_{n1}^\star[, p_n^\star\},$$
$$x_2^\star = \{[a_{12}^\star, b_{12}^\star[, p_1^\star; ...; [a_{i2}^\star, b_{i2}^\star[, p_i^\star; ...; [a_{n1}^\star, b_{n1}^\star[, p_n^\star\},$$

with a harmonized set of cumulative probabilities $\mathcal{W}^\star = \{w_0^\star, .., w_i^\star, .., w_n^\star\}$ and with harmonized quantile functions $Q_{x_1^\star}$ and $Q_{x_2^\star}$, the addition of x_1 with x_2 is obtained by the addition of $Q_{x_1^\star}$ with $Q_{x_2^\star}$, such that:

$$Q_{x_1+x_2}(p)= \begin{cases} a_{11}^\star + a_{12}^\star + \frac{p}{w_1^\star}(b_{11}^\star - a_{11}^\star + (b_{12}^\star - a_{12}^\star)), & 0 \leq p < w_1^\star \\ a_{21}^\star + a_{22}^\star + \frac{p-w_1^\star}{w_2^\star - w_1^\star}(b_{21}^\star - a_{21}^\star + (b_{22}^\star - a_{22}^\star)), & w_1^\star \leq p < w_2^\star \\ \vdots & \\ a_{i1}^\star + a_{i2}^\star + \frac{p-w_{i-1}^\star}{w_i^\star - w_{i-1}^\star}(b_{i1}^\star - a_{i1}^\star + (b_{i2}^\star - a_{i2}^\star)), & w_{i-1}^\star \leq p < w_i^\star \\ \vdots & \\ a_{n1}^\star + a_{n2}^\star + \frac{p-w_{n-1}^\star}{1-w_{n-1}^\star}(b_{n1}^\star - a_{n1}^\star + (b_{n2}^\star - a_{n2}^\star)), & w_{n-1}^\star \leq p \leq 1 \end{cases}. \tag{3.8}$$

If the notation with the centers and ranges of the subintervals ($c_{ij}^\star$ and $r_{ij}^\star$) is used, instead of the harmonized quantile functions being expressed as a function of the bounds of the subintervals ($a_{ij}^\star$ and $b_{ij}^\star$), the sum of x_1 with x_2 would be given by the following function:

$$Q_{x_1+x_2}(p)= \begin{cases} c_{11}^\star + c_{12}^\star + (\frac{2p}{w_1^\star} - 1)(\frac{r_{11}^\star + r_{12}^\star}{2}), & 0 \leq p < w_1^\star \\ c_{21}^\star + c_{22}^\star + (\frac{2(p-w_1^\star)}{w_2^\star - w_1^\star} - 1)(\frac{r_{21}^\star + r_{22}^\star}{2}), & w_1^\star \leq p < w_2^\star \\ \vdots & \\ c_{i1}^\star + c_{i2}^\star + (\frac{2(p-w_{i-1}^\star)}{w_i^\star - w_{i-1}^\star} - 1)(\frac{r_{i1}^\star + r_{i2}^\star}{2}), & w_{i-1}^\star \leq p < w_i^\star \\ \vdots & \\ c_{n1}^\star + c_{n2}^\star + (\frac{2(p-w_{n-1}^\star)}{1-w_{n-1}^\star} - 1)(\frac{r_{n1}^\star + r_{n2}^\star}{2}), & w_{n-1}^\star \leq p \leq 1 \end{cases}. \tag{3.9}$$

Through the analysis of the previous expression (3.9), it is easy to infer that the centers of the subintervals from the resulting histogram $x_1 + x_2$ correspond to the sum of the centers from the subintervals of $x_1^\star$ and $x_2^\star$, which are defined over the same range $[w_{i-1}^\star, w_i^\star[$ in the harmonized quantile functions. The same applies to the ranges of the subintervals: the resulting ranges of the subintervals of $x_1 + x_2$ match the sum of the ranges of the subintervals of $x_1^\star$ and $x_2^\star$. Hence, for each ordered subinterval i of the histogram originated from the sum of the harmonized histograms $x_1^\star$ and $x_2^\star$, its respective centers and ranges (c_i and r_i) are given by the following expressions:

$$\begin{aligned} c_i &= c_{i1}^\star + c_{i2}^\star, \\ r_i &= r_{i1}^\star + r_{i2}^\star, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{3.10}$$

If we want 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. This procedure can be thought of as the addition of a regular histogram x_1 with another histogram x_2, which is composed of only one subinterval [β, β] with an associated probability $p = 1$. Therefore, the set of cumulative probabilities of x_2, $\mathcal{W}_2$, would only have two elements: $w_{02} = 0$ and $w_{12} = 1$. If a harmonization procedure is performed on x_2 together with another regular histogram x_1 with n subintervals and a set $\mathcal{W}_1 = \{w_{01}, w_{11}, .., w_{n1}\}$, the harmonized version of the

histogram x_1 would remain the same ($x_1^\star$=x_1, since the harmonized set $\mathcal{W}^\star$ would be equal to $\mathcal{W}_1$) and x_2 would be transformed into the following harmonized histogram $x_2^\star$:

$$Q_{x_2^\star}(p)= \begin{cases} \beta, & 0 \leq p < w_{11} \\ \beta, & w_{11} \leq p < w_{21} \\ \vdots & \\ \beta, & w_{i-1\,1} \leq p < w_{i1} \\ \vdots & \\ \beta, & w_{n-1\,1} \leq p < 1 \end{cases}.$$

That is why the operation x_1 + β can be performed by simply adding β to the quantile function Q_{x_1} and no transformations related to the harmonization process need to be applied. Accordingly, when using the notation with the bounds of the subintervals, we have

$$Q_{x_1+\beta} = \begin{cases} a_{11} + \beta + \frac{p}{w_{11}}(b_{11} - a_{11}), & 0 \leq p < w_{11} \\ a_{21} + \beta + \frac{p-w_{11}}{w_{21}-w_{11}}(b_{21} - a_{21}), & w_{11} \leq p < w_{21} \\ \vdots & \\ a_{i1} + \beta + \frac{p-w_{i-1\,1}}{w_{i1}-w_{i-1\,1}}(b_{i1} - a_{i1}), & w_{i-1\,1} \leq p < w_{ij} \\ \vdots & \\ a_{n1} + \beta + \frac{p-w_{n-1\,1}}{1-w_{n-1\,1}}(b_{n1} - a_{n1}), & w_{n-1\,1} \leq p \leq 1 \end{cases}. \tag{3.11}$$

And with the notation of the centers and ranges of the subintervals:

$$Q_{x_1+\beta} = \begin{cases} c_{11} + \beta + (\frac{2p}{w_{11}} - 1)\frac{r_{11}}{2}, & 0 \leq p < w_{11} \\ c_{21} + \beta + (\frac{2(p-w_{11})}{w_{21}-w_{11}} - 1)\frac{r_{21}}{2}, & w_{11} \leq p < w_{21} \\ \vdots & \\ c_{i1} + \beta + (\frac{2(p-w_{i-1\,1})}{w_{i1}-w_{i-1\,1}} - 1)\frac{r_{i1}}{2}, & w_{i-1\,1} \leq p < w_{i1} \\ \vdots & \\ c_{n1} + \beta + (\frac{2(p-w_{n-1\,1})}{1-w_{n-1\,1}} - 1)\frac{r_{n1}}{2}, & w_{n-1\,1} \leq p \leq 1 \end{cases}. \tag{3.12}$$

From the previous expression, it is easily ascertained that the addition of a histogram x_1 with a constant $\beta \in \mathbb{R}$ only affects the centers of the subintervals of x_1 and the ranges suffer no alterations. For that reason, this operation corresponds to a translation of the subintervals to the left, if $\beta < 0$, and to a translation to the right, if $\beta > 0$. The resulting centers and ranges for the subinterval i of the histogram resulting from the operation $x_1 + \beta$ (c_i and r_i) are then given by the expressions:

$$\begin{aligned} c_i &= c_{i1} + \beta, \\ r_i &= r_{i1}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{3.13}$$

The operations corresponding to the sum of two histograms and the sum of a histogram with a

constant are illustrated in the example that follows.

Example 3.2.4. *Using the previously defined expression for the addition in (3.11), now we want to perform the sum of two histograms x_1 and x_2. To do this, we start by performing a harmonization on these two histograms, originally represented in Example 3.2.1, whereas Example 3.2.3 depicts the completed harmonization process. The resulting harmonized quantile functions $Q_{x_1^\star}$ and $Q_{x_2^\star}$, under the notation of the bounds of the subintervals, are given by*

$$Q_{x_1^\star}(p)= \begin{cases} -5 + \frac{p}{0.4} \times \frac{12}{5}, & 0 \leq p < 0.4 \\ -\frac{13}{5} + \frac{p-0.4}{0.1} \times \frac{3}{5}, & 0.4 \leq p < 0.5 \\ -2 + \frac{p-0.5}{0.3} \times 5, & 0.5 \leq p < 0.8 \\ 3 + \frac{p-0.8}{0.2}, & 0.8 \leq p \leq 1 \end{cases},$$

$$Q_{x_2^\star}(p) = \begin{cases} 3 + \frac{p}{0.4} \times 7, & 0 \leq p < 0.4 \\ 10 + \frac{p-0.4}{0.1} \times \frac{1}{3}, & 0.4 \leq p < 0.5 \\ \frac{31}{3} + \frac{p-0.5}{0.3}, & 0.5 \leq p < 0.8 \\ \frac{34}{3} + \frac{p-0.8}{0.2} \times \frac{2}{3}, & 0.8 \leq p \leq 1 \end{cases}.$$

To compute $x_1 + x_2$, we just need to add these two harmonized quantile functions, which results in the following quantile function $Q_{x_1+x_2}$:

$$Q_{x_1+x_2}= \begin{cases} -2 + \frac{p}{0.4} \times \frac{47}{5}, & 0 \leq p < 0.4 \\ \frac{37}{5} + \frac{p-0.4}{0.1} \times \frac{14}{15}, & 0.4 \leq p < 0.5 \\ \frac{25}{3} + \frac{p-0.5}{0.3} \times 6, & 0.5 \leq p < 0.8 \\ \frac{43}{3} + \frac{p-0.8}{0.2} \times \frac{5}{3}, & 0.8 \leq p \leq 1 \end{cases}.$$

The harmonized histograms $x_1^\star$ and $x_2^\star$, as well as the resulting histogram from their addition are represented in Figure 3.3, where it can be observed that the ranges and centers of the subintervals of $x_1 + x_2$ were obtained from the sum of the centers and ranges of the subintervals of $x_1^\star$ and $x_2^\star$.

If we want to make the operation x_1+5, it is not necessary to perform the harmonization on histogram x_1. We can use its original quantile function Q_{x_1}:

$$Q_{x_1}(p)= \begin{cases} -5 + \frac{p}{0.5} \times 3, & 0 \leq p < 0.5 \\ -2 + \frac{p-0.2}{0.3} \times 5, & 0.5 \leq p < 0.8 \\ 3 + \frac{p-0.8}{0.2}, & 0.8 \leq p \leq 1 \end{cases},$$

and simply add 5 to Q_{x_1}, which gives

$$Q_{x_1+5}(p)= \begin{cases} 0 + \frac{p}{0.5} \times 3, & 0 \leq p < 0.5 \\ 3 + \frac{p-0.2}{0.3} \times 5, & 0.5 \leq p < 0.8 \\ 8 + \frac{p-0.8}{0.2}, & 0.8 \leq p \leq 1 \end{cases}.$$

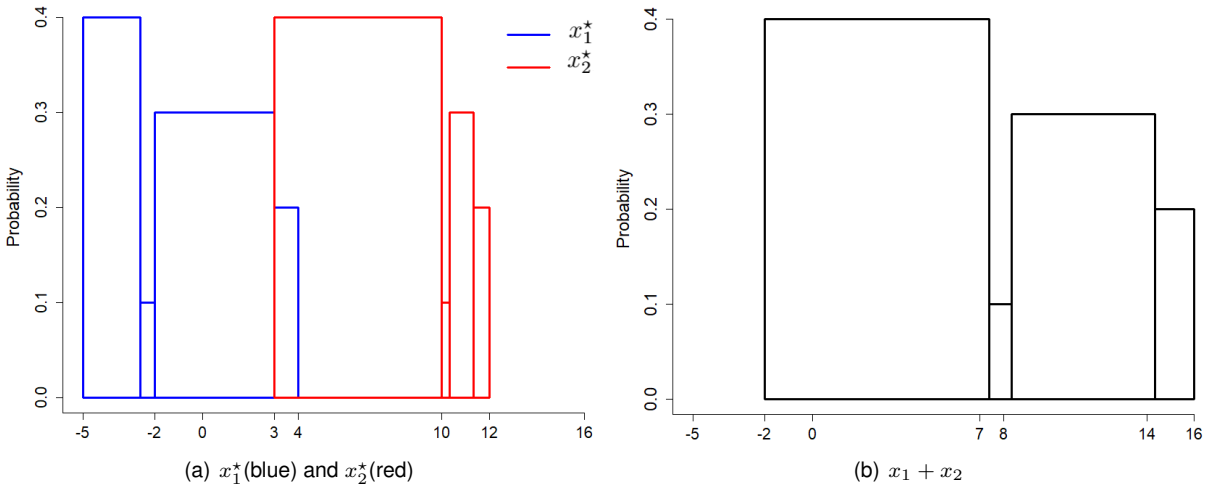


Figure 3.3: The histogram $x_1^\star$ and $x_2^\star$ and the histogram resulting from $x_1 + x_2$ from Example 3.2.4.

Figure 3.4 displays the original histogram x_1 and the histogram obtained from the operation $x_1 + 5$. There, it can be observed that the histogram $x_1 + 5$ corresponds simply to a translation to the right of 5 units of the histogram x_1.

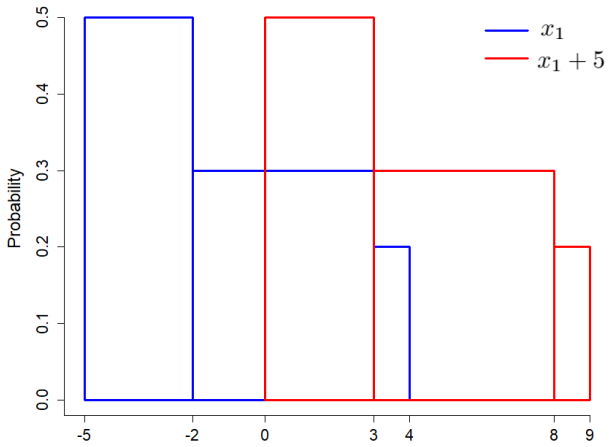


Figure 3.4: Original histogram x_1 (in blue) and the histogram $x_1 + 5$ (in red) from Example 3.2.4.

□

Another arithmetic operation that is important to define for quantile functions is the multiplication by a constant $\beta \in \mathbb{R}$. It is necessary to be careful when specifying this operation, because it can lead to unpractical results if it is considered as the straightforward product of the constant β with every branch of the quantile function, as subsequently shown.

The product of a constant $\beta \in \mathbb{R}$ with a quantile function of a histogram x_1 with n_1 subintervals, Q_{x_1}, (written under the notation with the bounds of the subintervals), produces the following function:

$$\beta Q_{x_1} = \begin{cases} \beta \times a_{11} + \frac{p}{w_{11}}(b_{11} - a_{11})\beta, & 0 \leq p < w_{11} \\ \beta \times a_{21} + \frac{p - w_{11}}{w_{21}}(b_{21} - a_{21})\beta, & w_{11} \leq p < w_{21} \\ \vdots & \\ \beta \times a_{i1} + \frac{p - w_{i-1\,1}}{w_{i1} - w_{i-1\,1}}(b_{i1} - a_{i1})\beta, & w_{i-1\,1} \leq p < w_{i1} \\ \vdots & \\ \beta \times a_{n_1 1} + \frac{p - w_{n_1 - 1\,1}}{1 - w_{n_1 - 1\,1}}(b_{n_1 1} - a_{n_1 1})\beta, & w_{n_1 - 1\,1} \leq p \leq 1 \end{cases}.$$

When the constant β is positive, there are no issues with this approach. In that case, the resulting function from βQ_{x_1} would still be a non-decreasing function. However, when $\beta < 0$, the same does not occur and the outcome is a decreasing function. As it was mentioned before, one of the properties of the quantile functions is that they must always be non-decreasing. Consequently, this function generated from the operation βQ_{x_1}, with $\beta < 0$, cannot be considered to be a quantile function and, subsequently, cannot be representative of an observation of a histogram variable. Therefore, a solution must be found to overcome this problem, such that the multiplication of a quantile function by a constant $\beta \in \mathbb{R}$ always produces a non-decreasing function $\forall \beta \in \mathbb{R}$.

In Example 3.2.5, it is demonstrated why, using this method, the computation of the product of a constant with a histogram is not valid.

Example 3.2.5. *Considering a histogram x_1 with the expression*

$$x_1 = \{[0, 4[, 0.6; [4, 10[, 0.2; [10, 14], 0.2\},$$

and, using the notation with the bounds of the subintervals, with the quantile function Q_{x_1}:

$$Q_{x_1}(p) = \begin{cases} \frac{p}{0.6} \times 4, & 0 \leq p < 0.6 \\ 4 + \frac{p - 0.6}{0.2} \times 6, & 0.6 \leq p < 0.8 \\ 10 + \frac{p - 0.8}{0.2} \times 4, & 0.8 \leq p \leq 1 \end{cases}.$$

According to the previous definition of the operation βQ_{x_1}, as the product of $\beta = -1$ with every branch of Q_{x_1}, the calculation of $-Q_{x_1}$ generates the following function:

$$-Q_{x_1}(p) = \begin{cases} -\frac{p}{0.6} \times 4, & 0 \leq p < 0.6 \\ -4 - \frac{p - 0.6}{0.2} \times 6, & 0.6 \leq p < 0.8 \\ -10 - \frac{p - 0.8}{0.2} \times 4, & 0.8 \leq p \leq 1 \end{cases}.$$

The graphic representation of $-Q_{x_1}(p)$ is displayed in Figure 3.5. This function is clearly decreasing and, consequently, it is not a quantile function. The equivalent "histogram" $-x_1$ of this function has the expression

$$-x_1 = \{[0, -4[, 0.6; [-4, -10[, 0.2; [-10, -14], 0.2\}$$

The previous expression of $-x_1$ cannot be representative of a histogram: the value of the lower bounds of its subintervals are higher than the value of the upper bounds ($a_{i1} > b_{i1}$). This does not

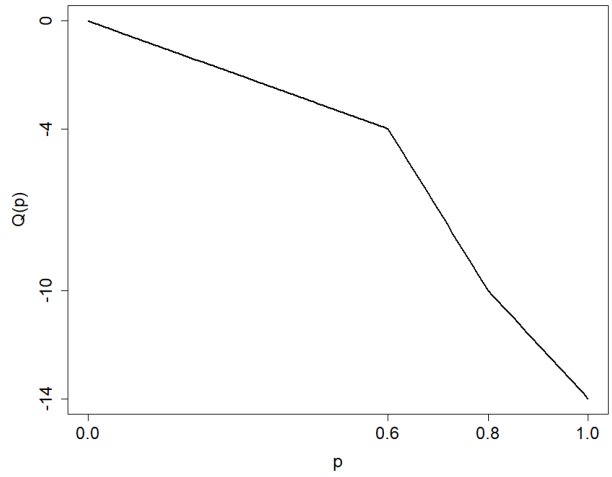


Figure 3.5: Graphic representation of $-Q_{x1}(p)$ from the Example 3.2.5.

respect the condition for the subintervals of histograms, which states that $a_{ij} \leq b_{ij}$, $\forall i \in \{1, .., n_j\}$. *Hence, another method must be used in order to enable us to perform the multiplication of quantile functions by negative constants.* □

The most correct method to carry out the operation βx_1, when $\beta < 0$, is firstly to do the transformation $Q_{x_1}(1-p)$ on the quantile function and then multiply the resulting function by β. With this procedure, the resulting function is always non-decreasing, and therefore it belongs to the class of quantile functions. Furthermore, 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. Therefore, the transformation $-Q_{x_1}(1-p)$ not only solves the issue with the previous definition of $-Q_{x_1}$, but it also represents the result in the form of the symmetric of the original histogram, as desired.

Consequently, the multiplication of a histogram x_1 by a real number β will be split into two cases: one for the instances when $\beta \geq 0$, where a direct multiplication of β by Q_{x_1} is done, and another for the cases $\beta < 0$, where the previously mentioned transformation $\beta Q_{x_1}(1-p)$ is applied beforehand. Accordingly, βQ_{x_1} is calculated in the following manner, with the notation that uses the bounds of the subintervals:

- for $\beta \geq 0$:

$$\beta Q_{x_1} = \begin{cases} \beta \times a_{11} + \frac{p}{w_{11}}(b_{11} - a_{11})\beta, & 0 \leq p < w_{11} \\ \beta \times a_{21} + \frac{p - w_{11}}{w_{21} - w_{11}}(b_{21} - a_{21})\beta, & w_{11} \leq p < w_{21} \\ \vdots & \\ \beta \times a_{i1} + \frac{p - w_{i-1\,1}}{w_{i1} - w_{i-1\,1}}(b_{i1} - a_{i1})\beta, & w_{i-1\,1} \leq p < w_{i1} \\ \vdots & \\ \beta \times a_{n_1 1} + \frac{p - w_{n_1-1\,1}}{1 - w_{n_1-1\,1}}(b_{n_1 1} - a_{n_1 1})\beta, & w_{n_1-1\,1} \leq p \leq 1 \end{cases}; \tag{3.14}$$

- for $\beta < 0$:

$$\beta Q_{x_1}(1-p) = \begin{cases} \beta \times a_{11} + \frac{(1-p)}{w_{11}}(b_{11}-a_{11})\beta, & 0 \leq 1-p < w_{11} \\ \beta \times a_{21} + \frac{(1-p)-w_{11}}{w_{21}-w_{11}}(b_{21}-a_{21})\beta, & w_{11} \leq 1-p < w_{21} \\ \vdots & \\ \beta \times a_{i1} + \frac{(1-p)-w_{i-1\,1}}{w_{i1}-w_{i-1\,1}}(b_{i1}-a_{i1})\beta, & w_{i-1\,1} \leq 1-p < w_{i1} \\ \vdots & \\ \beta \times a_{n_1 1} + \frac{(1-p)-w_{n_1-1\,1}}{1-w_{n_1-1\,1}}(b_{n_1 1}-a_{n_1 1})\beta, & w_{n_1-1\,1} \leq 1-p \leq 1 \end{cases} =$$

$$= \begin{cases} \beta \times b_{n_1 1} + \frac{p}{w^{\diamond}_{11}}(b_{n_1 1}-a_{n_1 1})|\beta|, & 0 \leq p < w^{\diamond}_{11} \\ \beta \times b_{n_1-1\,1} + \frac{p-w^{\diamond}_{11}}{w^{\diamond}_{21}-w^{\diamond}_{11}}(b_{n_1-1\,1}-a_{n_1-1\,1})|\beta|, & w^{\diamond}_{11} \leq p < w^{\diamond}_{21} \\ \vdots & \\ \beta \times b_{i1} + \frac{p-w^{\diamond}_{i-1\,1}}{w^{\diamond}_{i1}-w^{\diamond}_{i-1\,1}}(b_{i1}-a_{i1})|\beta|, & w^{\diamond}_{i-1\,1} \leq p < w^{\diamond}_{i1} \\ \vdots & \\ \beta \times b_{11} + \frac{p-w^{\diamond}_{n_1-1\,1}}{1-w^{\diamond}_{n_1-1\,1}}(b_{11}-a_{11})|\beta|, & w^{\diamond}_{n_1-1\,1} \leq p \leq 1 \end{cases}, \tag{3.15}$$

where $\mathcal{W}_1^{\diamond} = \{w_{01}^{\diamond}, w_{11}^{\diamond}, .., w_{n_1 1}^{\diamond}\}$ is the new set of cumulative probabilities for the cases when the order of the subintervals is reversed, due to the transformation $Q_{x_1}(1-p)$. These $w_{i1}^{\diamond}$ can be calculated using the original set $\mathcal{W}_1$ through the expression: $w_{i1}^{\diamond} = 1 - w_{n_1-i+1}$. Accordingly, it is possible to verify in the previous expressions that, for the case where $\beta < 0$, there occurs an inversion of the order of the original subintervals. The first subinterval from the initial histogram with an associated probability p_{11} is the last in the new histogram with the same associated probability, and so on. Therefore, $-Q_{x_1}(1-p)$ is indeed the inverse of Q_{x_1}.

When the notation with the centers and ranges of the subintervals is used, the resulting function is

- for $\beta \geq 0$:

$$\beta Q_{x_1} = \begin{cases} \beta \times c_{11} + (\frac{2p}{w_{11}} - 1)\frac{r_{11}}{2} \times \beta, & 0 \leq p < w_{11} \\ \beta \times c_{21} + (\frac{2(p-w_{11})}{w_{21}-w_{11}} - 1)\frac{r_{21}}{2} \times \beta, & w_{11} \leq p < w_{21} \\ \vdots & \\ \beta \times c_{i1} + (\frac{2(p-w_{i-1\,1})}{w_{i1}-w_{i-1\,1}} - 1)\frac{r_{i1}}{2} \times \beta, & w_{i-1\,1} \leq p < w_{i1} \\ \vdots & \\ \beta \times c_{n_1 1} + (\frac{2(p-w_{n_1-1\,1})}{1-w_{n_1-1\,1}} - 1)\frac{r_{n_1 1}}{2} \times \beta, & w_{n_1-1\,1} \leq p \leq 1 \end{cases}; \tag{3.16}$$

- for $\beta < 0$:

$$\beta Q_{x_1}(1-p) = \begin{cases} \beta \times c_{n_1 1} + (\frac{2p}{w_{11}^{\diamond}} - 1)\frac{r_{n_1 1}}{2} \times |\beta|, & 0 \leq p < w_{11}^{\diamond} \\ \beta \times c_{n_1-1\,1} + (\frac{2(p-w_{11}^{\diamond})}{w_{21}^{\diamond} - w_{11\diamond}} - 1)\frac{r_{n_1-1\,1}}{2} \times |\beta|, & w_{11}^{\diamond} \leq p < w_{21}^{\diamond} \\ \vdots & \\ \beta \times c_{i1} + (\frac{2(p-w_{i-1\,1}^{\diamond})}{w_{i1}^{\diamond} - w_{i-1\,1}^{\diamond}} - 1)\frac{r_{i1}}{2} \times |\beta|, & w_{i-1\,1}^{\diamond} \leq p < w_{i1}^{\diamond} \\ \vdots & \\ \beta \times c_{11} + (\frac{2(p-w_{n_1-1\,1}^{\diamond})}{1 - w_{n_1-1\,1\diamond}} - 1)\frac{r_{11}}{2} \times |\beta|, & w_{n_1-1\,1}^{\diamond} \leq p \leq 1 \end{cases}. \tag{3.17}$$

The analysis of the previous expressions for the centers and ranges enables us to conclude that the result of 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 $\beta > 0$:

$$\begin{aligned} c_i &= \beta c_{i1}, \\ r_i &= \beta r_{i1}, \quad \text{with } i \in \{1, .., n\}, \end{aligned} \tag{3.18}$$

- for $\beta < 0$:

$$\begin{aligned} c_i &= \beta c_{n_1 - i + 1\,1}, \\ r_i &= |\beta| r_{n_1 - i + 1\,1}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{3.19}$$

Hence, it is guaranteed that the ranges of the resulting subintervals are always non-negative. It is also possible to conclude that, apart from the possible inversion of the order of the subintervals that occurs when $\beta < 0$, these operations follow the same reasoning as in Moore's Interval Algebra for each subinterval of the histograms.

The next arithmetic operation that is important to define is the difference between two histograms x_1 and x_2: $x_1 - x_2$. 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 first on x_2, and $x_2^{\star}$ is obtained, followed by the calculation of its inverse $-(x_2^{\star})$, it will not be as efficient. This is due to the fact that the branches of the quantile functions of x_1 and $-(x_2^{\star})$ could not be defined over the same subintervals of cumulative probabilities, which would require another harmonization to be performed, 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 following notation for the bounds of the subintervals:

$$(-x_j)^{\star} = \{[-b_{nj}^{\diamond\star}, -a_{nj}^{\diamond\star}[, p_{nj}^{\diamond\star}; ..; [-b_{1j}^{\diamond\star}, -a_{1j}^{\diamond\star}], p_{1j}^{\diamond\star}\}, \tag{3.20}$$

and for the notation with a set of centers and ranges:

$$(-x_j)^\star = \{(-c_{nj}^{\diamond\star}, r_{nj}^{\diamond\star}), p_{nj}^{\diamond\star}; ..; (-c_{1j}^{\diamond\star}, r_{1j}^{\diamond\star}), p_{1j}^{\diamond\star}\}, \tag{3.21}$$

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 first and only afterwards was the harmonization performed. Furthermore, the subinterval i of the resulting histogram of this operation is represented with the index $n-i+1$ in this notation, to relate the upcoming operations with the original order of the subintervals and, this way, better understand the resulting centers and ranges. However, the harmonized set of cumulative probablities that results from the harmonization procedure of the sets W_1 and $\mathcal{W}_2^{\diamond}$ is still represented by $\mathcal{W}^\star$. As a result, it is possible to calculate the result of the difference operation $x_1 - x_2$, under the notation of the bounds of the subintervals, using the following expression:

$$Q_{x_1-x_2}(p) = \begin{cases} a_{11}^\star - b_{n2}^{\diamond\star} + \frac{p}{w_1^\star}(b_{11}^\star - a_{11}^\star + (b_{n2}^{\diamond\star} - a_{n2}^{\diamond\star})), & 0 \leq p < w_1^\star \\ a_{21}^\star - b_{n-1\,2}^{\diamond\star} + \frac{p-w_1^\star}{w_2^\star - w_1^\star}(b_{21}^\star - a_{21}^\star + (b_{n-1\,2}^{\diamond\star} - a_{n-1\,2}^{\diamond\star})), & w_1^\star \leq p < w_2^\star \\ \vdots & \\ a_{i1}^\star - b_{i2}^{\diamond\star} + \frac{p-w_{i-1}^\star}{w_i^\star - w_{i-1}^\star}(b_{i1}^\star - a_{i1}^\star + (b_{i2}^{\diamond\star} - a_{i2}^{\diamond\star})), & w_{i-1}^\star \leq p < w_i^\star \\ \vdots & \\ a_{n1}^\star - b_{12}^{\diamond\star} + \frac{p-w_{n-1}^\star}{1-w_{n-1}^\star}(b_{n1}^\star - a_{n1}^\star + (b_{12}^{\diamond\star} - a_{12}^{\diamond\star})), & w_{n-1}^\star \leq p \leq 1 \end{cases}, \tag{3.22}$$

and for the centers and ranges notation:

$$Q_{x_1-x_2}(p) = \begin{cases} c_{11}^\star - c_{n2}^{\diamond\star} + (\frac{2p}{w_{11}^\star} - 1)(\frac{r_{11}^\star + r_{n2}^{\diamond\star}}{2}), & 0 \leq p < w_1^\star \\ c_{21}^\star - c_{n-1\,2}^{\diamond\star} + (\frac{2(p-w_{11}^\star)}{w_{21}^\star - w_{11}^\star} - 1)(\frac{r_{21}^\star + r_{n-1\,2}^{\diamond\star}}{2}), & w_1^\star \leq p < w_2^\star \\ \vdots & \\ c_{i1}^\star - c_{i2}^{\diamond\star} + (\frac{2(p-w_{i-1\,1}^\star)}{w_{i1}^\star - w_{i-1\,1}^\star} - 1)(\frac{r_{i1}^\star + r_{i2}^{\diamond\star}}{2}), & w_{i-1}^\star \leq p < w_i^\star \\ \vdots & \\ c_{n1}^\star - c_{12}^{\diamond\star} + (\frac{2(p-w_{n_1-1\,1}^\star)}{1-w_{n_1-1\,1}^\star} - 1)(\frac{r_{n1}^\star + r_{12}^{\diamond\star}}{2}), & w_{n-1}^\star \leq p \leq 1 \end{cases}. \tag{3.23}$$

The order of the subintervals is reversed in the histogram $(-x_2)^\star$, hence the centers and ranges for the resulting subinterval i with $i = \{1, .., n\}$ of $x_1 - x_2$ are given by

$$\begin{aligned} c_i &= c_{i1}^\star - c_{n-i+1\,2}^{\diamond\star}, \\ r_i &= r_{i1}^\star + r_{n-i+1\,2}^{\diamond\star}, \quad \text{with } i \in \{1, .., n\}. \end{aligned} \tag{3.24}$$

Thus, the resulting centers of the subinterval i of $x_1 - x_2$ are obtained from the difference between the centers of the subinterval i of $x_1^\star$ and the subinterval $n - i + 1$ of $(-x_2)^\star$, which had the order of the

indices of its subintervals reversed. On the other hand, the ranges of the subintervals of $x_1 - x_2$ are obtained from the addition of the ranges of those two subintervals. Therefore, the subtraction operation of two histograms also follows a logic similar to the one presented in Moore's Interval Algrebra for each of their subintervals.

The necessary steps to compute the subtraction of two histograms are shown in Example 3.2.6.

Example 3.2.6. *Considering the histogram x_1, described by the expression*

$$x_1 = \{[0, 4[, 0.6; [4, 10[, 0.2; [10, 14], 0.2\},$$

and with the respective quantile function Q_{x_1}:

$$Q_{x_1}(p) = \begin{cases} \frac{p}{0.6} \times 4, & 0 \leq p < 0.6 \\ 4 + \frac{p-0.6}{0.2} \times 6, & 0.6 \leq p < 0.8 \\ 10 + \frac{p-0.8}{0.2} \times 4, & 0.8 \leq p \leq 1 \end{cases}.$$

When $-x_1$ is obtained according to (3.15), it was previously determined that the resulting histogram corresponds to the inverse of x_1 along the y-axis. Consequently, the expressions for $-x_1$ and its quantile function are

$$-x_1 = \{[-14, -10[, 0.2; [-10, -4[, 0.2; [-4, 0], 0.6\},$$

$$Q_{-x_1}(p) = \begin{cases} -14 + \frac{p}{0.2} \times 4, & 0 \leq p < 0.2 \\ -10 + \frac{p-0.2}{0.2} \times 6, & 0.2 \leq p < 0.4 \\ -4 + \frac{p-0.4}{0.6} \times 4, & 0.4 \leq p \leq 1 \end{cases}.$$

The comparison between the original histogram x_1 and its inverse $-x_1$ is displayed in Figure 3.6, where it is possible to verify that the outcome of the operation βx_1 with $\beta = -1$ is the symmetric of x_1.

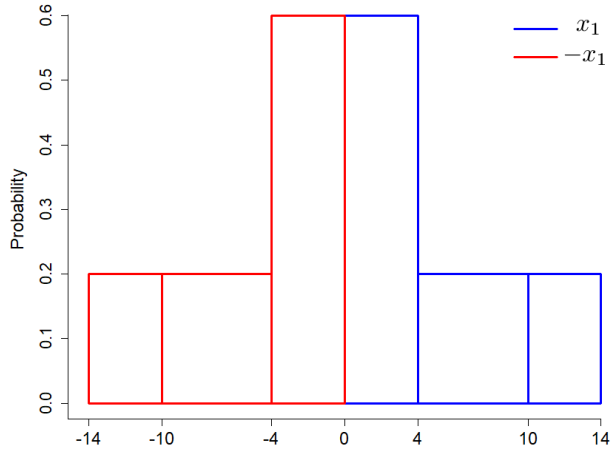


Figure 3.6: Original histogram x_1 and its symmetric $-x_1$ from Example 3.2.6.

To calculate $x_1 - x_1$ according to (3.20), first it is necessary to harmonize the histograms x_1 and $-x_1$ to fit the set of harmonized cumulative probabilities $\mathcal{W}^\star = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. This generates the following expressions for the harmonized quantile functions of $x_1^\star$ and (-x_1)*, respectively:

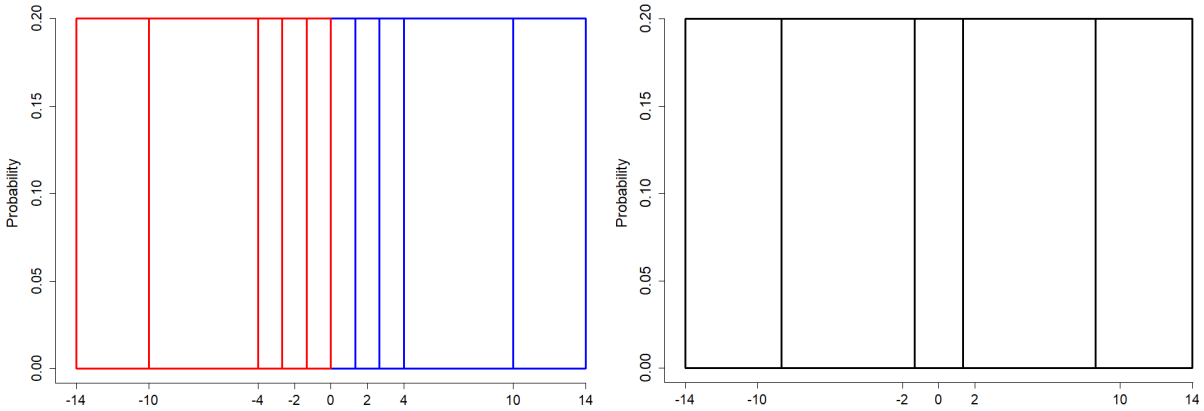
$$Q_{x_1^\star}(p) = \begin{cases} \frac{p}{0.2} \times \frac{4}{3}, & 0 \leq p < 0.2 \\ \frac{4}{3} + \frac{p-0.2}{0.2} \times \frac{4}{3}, & 0.2 \leq p < 0.4 \\ \frac{8}{3} + \frac{p-0.4}{0.2} \times \frac{4}{3}, & 0.4 \leq p < 0.6 \\ 4 + \frac{p-0.6}{0.2} \times 6, & 0.6 \leq p < 0.8 \\ 10 + \frac{p-0.8}{0.2} \times 4, & 0.8 \leq p \leq 1 \end{cases},$$

$$Q_{(-x_1)^\star}(p) = \begin{cases} -14 + \frac{p}{0.2} \times 4, & 0 \leq p < 0.2 \\ -10 + \frac{p-0.2}{0.2} \times 6, & 0.2 \leq p < 0.4 \\ -4 + \frac{p-0.4}{0.2} \times \frac{4}{3}, & 0.4 \leq p < 0.6 \\ -\frac{8}{3} + \frac{p-0.2}{0.3} \times \frac{4}{3}, & 0.6 \leq p < 0.8 \\ -\frac{4}{3} + \frac{p-0.8}{0.2} \times \frac{4}{3}, & 0.8 \leq p \leq 1 \end{cases}.$$

Now, by doing a regular addition of these two quantile functions $Q_{x_1^\star}(p)$ and $Q_{(-x_1)^\star}(p)$, the following quantile function representative of the difference $x_1 - x_1$ is obtained:

$$Q_{x_1 - x_1}(p) = \begin{cases} -14 + \frac{p}{0.2} \times \frac{16}{3}, & 0 \leq p < 0.2 \\ -\frac{26}{3} + \frac{p-0.2}{0.2} \times \frac{22}{3}, & 0.2 \leq p < 0.4 \\ -\frac{4}{3} + \frac{p-0.4}{0.2} \times \frac{8}{3}, & 0.4 \leq p < 0.6 \\ \frac{4}{3} + \frac{p-0.2}{0.3} \times \frac{22}{3}, & 0.6 \leq p < 0.8 \\ \frac{26}{3} + \frac{p-0.8}{0.2} \times \frac{16}{3}, & 0.8 \leq p \leq 1 \end{cases}.$$

The harmonized histograms $x_1^\star$ and $(-x_1)^\star$ and the histogram that is the result of the difference $x_1 - x_1$ are represented in 3.7. There, it is possible to observe that the histogram $x_1 - x_1$ is symmetric with respect to the y-axis.



(a) $x_1^\star$ (blue) and $(-x_1)^\star$ (red) (b) $x_1 - x_1$

Figure 3.7: The harmonized histogram $x_1^\star$ and $(-x_1)^\star$ and the histogram $x_1 - x_1$ from Example 3.2.6.

□

As it can be observed in the previous example, it is also interesting to mention that the histograms generated by the difference operation between two equal histograms, $x_1 - x_1$ are always symmetric

with respect to the y-axis, and not the null function, as it would be expected. However, it will be seen in Chapter 4 that the histograms with this characteristic (symmetric with respect to the y-axis) have a symbolic mean equal to 0.

3.2.3 Histogram algebra

The first objective of our work is to create an algebra that sums up, in a simple and general manner, 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 (3.10), (3.13), (3.18), (3.19), and (3.20) of the previous sub-chapter, regarding the resulting centers and ranges for each subinterval i of the histogram generated for each operation. These results are summarized in Table 3.4.

Table 3.4: Centers and ranges of the subintervals resulting from the quantile operations.

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

Comparing these results with those obtained in Table 3.1 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. It was previously mentioned that this may be problematic, due to the loss of significance of these intervals/subintervals when the ranges become too large. Nonetheless, this is still the best option to perform arithmetic operations with intervals/subintervals, since, when interval algebras that allow the shrinking of the ranges in the operations are used (as it is the case of the Extended Interval Algebra), it can lead to even bigger issues as a result of the creation of degenerate intervals/subintervals.

When defining the operation βx_1 and the linear combinations of quantile functions $\beta_1 x_1 + \beta_2 x_2$, it is necessary to be careful because the order by which the harmonization is performed can lead to different results: $\beta_j \times (x_j)^{\star} \neq (\beta_j x_j)^{\star}$. It was previously observed that the harmonization should always be executed on the histograms $\beta_j x_j$, so as to achieve better results. However, this causes the reversion of the order of the original subintervals when $\beta < 0$. In order not to mix the cases where $\beta < 0$ and $\beta > 0$, the notation in the expressions (3.20) and (3.21) was created. The computation of $(\beta_j x_j)^{\star}$ is equivalent to $|\beta_j|(sign(\beta_j)x_j)^{\star}$, where $sign(\beta_j)$ is a function representative of the sign of each β_j, such that:

$$sign(\beta_j) = \begin{cases} -1, & \beta_j < 0 \\ 0, & \beta_j = 0 \\ 1, & \beta_j > 0 \end{cases}.$$

This allows the representation of the centers for each subinterval i that resulted from the operation $|\beta_j|(sign(\beta_j)x_j)^\star$ as

$$c_{ij} = \begin{cases} \beta_j c^\star_{ij}, & \beta_j \geq 0 \\ |\beta_j|(-c_{n-i+1\,j})^{\diamond\star}, & \beta_j < 0 \end{cases},$$

and for the ranges as

$$r_{ij} = \begin{cases} \beta_j r^\star_{ij}, & \beta_j \geq 0 \\ |\beta_j| r^{\diamond\star}_{n-i+1\,j}, & \beta_j < 0 \end{cases}.$$

Accordingly, 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. The values for the centers and ranges of the subinterval i of the histogram produced by the linear combination of two histograms $\beta_1 x_1 + \beta_2 x_2$ are displayed in Table 3.5.

Table 3.5: Different cases for the linear combination of two quantile functions.

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

By analyzing the previous table, one can conclude that it is 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, $c^\bullet_{ij}$ and $r^\bullet_{ij}$ are defined as follows:

$$c^\bullet_{ij} = \begin{cases} c^\star_{ij}, & \beta_j \geq 0 \\ -c^{\diamond\star}_{n-i+1\,j}, & \beta_j < 0 \end{cases}, \; r^\bullet_{ij} = \begin{cases} r^\star_{ij}, & \beta_j \geq 0 \\ r^{\diamond\star}_{n-i+1\,j}, & \beta_j < 0 \end{cases}, \tag{3.25}$$

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

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

$$p^\bullet_{ij} = \begin{cases} p^\star_{ij}, & \beta_j \geq 0 \\ p^{\diamond\star}_{n-i+1\,j}, & \beta_j < 0 \end{cases}. \tag{3.26}$$

Hence, using the notation in (3.25), a general formula for the resulting center and range of the subinterval $i \in \{1, .., n\}$ can be determined for the linear combination of two histograms $\beta_1 x_1 + \beta_2 x_2$, with $\beta_1, \beta_2 \in \mathbb{R}$:

$$\begin{aligned} c_i &= |\beta_1| c^\bullet_{i1} + |\beta_2| c^\bullet_{i2}, \\ r_i &= |\beta_1| r^\bullet_{i1} + |\beta_2| r^\bullet_{i2}. \end{aligned} \tag{3.27}$$

The other operations can also be altered according to this notation. Consequently, to account for these changes, Table 3.4 can be updated, as displayed in Table 3.6.

Table 3.6: Centers and ranges resulting from the quantile operations witn the notation in (3.25).

Operation	Center i	Range i
$x_1 + x_2$	$c^\bullet_{i1} + c^\bullet_{i2}$	$r^\bullet_{i1} + r^\bullet_{i2}$
$x_1 + \beta$	$c_{i1} + \beta$	r_{i1}
βx_1	$\lvert\beta\rvert c^\bullet_{i1}$	$\lvert\beta\rvert r^\bullet_{i1}$
$x_1 - x_2$	$c^\bullet_{i1} + c^\bullet_{i2}$	$r^\bullet_{i1} + r^\bullet_{i2}$
$\beta_1 x_1 + \beta_2 x_2$	$\lvert\beta_1\rvert c^\bullet_{i1} + \lvert\beta_2\rvert c^\bullet_{i2}$	$\lvert\beta_1\rvert r^\bullet_{i1} + \lvert\beta_2\rvert r^\bullet_{i2}$

Using the information available in Table 3.6, it is possible to deduce a general formula for the linear combination of histograms in this algebra. Hence, to do operations with k histograms x_j with $j = \{1, .., k\}$, which have n_j subintervals y_{ij}, $i = \{1, .., n_j\}$ with associated probabilities p_{ij} (before a harmonization is applied on them), and are represented as follows:

$$x_j = \{y_{1j}, p_{1j}; y_{2j}, p_{2j}; ..., y_{n_j j}, p_{n_j j}\}=$$
$$= \{c_{1j} \pm \tfrac{1}{2} r_{1j}, p_{1j}; c_{2j} \pm \tfrac{1}{2} r_{2j}, p_{2j}; ...; c_{n_j j} \pm \tfrac{1}{2} r_{n_j j}, p_{n_j j}\},$$

it is necessary to define the vector of constants $\boldsymbol{\beta}$ and the vector of histograms $\boldsymbol{x}$, such that:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}.$$

The next step is to perform a harmonization on all the k histograms contained in $\boldsymbol{x}$. This procedure depends on the corresponding values β_j in $\boldsymbol{\beta}$. As previously mentioned, it is necessary to first compute each $(sign(\beta_j)x_j)$, and only then harmonize the resulting histograms. As a result, k histograms $(sign(\beta_j x_j))^\star$, each one characterized by n subintervals, are obtained. The subintervals resulting from this process are denoted as $y^\bullet_{ij} = [c^\bullet_{ij} - \frac{1}{2} r^\bullet_{ij}, c^\bullet_{ij} + \frac{1}{2} r^\bullet_{ij}]$, according to the notation in (3.25). Therefore, it is also useful to specify the matrices $\boldsymbol{C}^\bullet$ and $\boldsymbol{R}^\bullet$ that contain all the $c^\bullet_{ij}$ and $r^\bullet_{ij}$ values:

$$\boldsymbol{C}^\bullet = \begin{bmatrix} c^\bullet_{11} & c^\bullet_{12} & \dots & c^\bullet_{1k} \\ c^\bullet_{21} & c^\bullet_{22} & \dots & c^\bullet_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c^\bullet_{i1} & c^\bullet_{i2} & \dots & c^\bullet_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ c^\bullet_{n1} & c^\bullet_{n2} & \dots & c^\bullet_{nk} \end{bmatrix}, \boldsymbol{R}^\bullet = \begin{bmatrix} r^\bullet_{11} & r^\bullet_{12} & \dots & r^\bullet_{1k} \\ r^\bullet_{21} & r^\bullet_{22} & \dots & r^\bullet_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r^\bullet_{i1} & r^\bullet_{i2} & \dots & r^\bullet_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ r^\bullet_{n1} & r^\bullet_{n2} & \dots & r^\bullet_{nk} \end{bmatrix}.$$

The rows of these two matrices are associated with the subinterval i, whereas the j-th column corresponds to the harmonized histogram j.

Before proceeding, there is one last matrix that needs to be defined: the matrix of the harmonized subinterval probabilities $\boldsymbol{P}^\bullet$. It is considered that the vector of the n harmonized probabilities associated to each histogram j is given by the vector $\boldsymbol{p}^\bullet_j = [p^\bullet_{1j}, .., p^\bullet_{ij}, .., p^\bullet_{nj}]$ (under the notation in (3.26)). Since

the harmonized probabilities are the same for all the histograms, the columns of the matrix $\boldsymbol{P}^\bullet$ are all the same and equal to $\boldsymbol{p}_j^\bullet = (p_1^\bullet, .., p_i^\bullet, .., p_n^\bullet)^t$. Therefore, we have

$$\boldsymbol{P}^\bullet = \begin{bmatrix} p_1^\bullet & p_1^\bullet & \dots & p_1^\bullet \\ p_2^\bullet & p_2^\bullet & \dots & p_2^\bullet \\ \vdots & \vdots & \ddots & \vdots \\ p_i^\bullet & p_i^\bullet & \dots & p_i^\bullet \\ \vdots & \vdots & \ddots & \vdots \\ p_n^\bullet & p_n^\bullet & \dots & p_n^\bullet \end{bmatrix}.$$

Having defined these vectors and matrices, all the conditions to determine a general expression for the linear combinations of histograms are met. Being $\boldsymbol{\beta}^t\boldsymbol{x}$ a linear combination of k histograms $\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$, its general expression is given by

$$\boldsymbol{\beta}^t\boldsymbol{x} = \{\boldsymbol{C}_1^\bullet|\boldsymbol{\beta}| \pm \frac{1}{2}\boldsymbol{R}_1^\bullet|\boldsymbol{\beta}|, p_1^\bullet; \boldsymbol{C}_2^\bullet|\boldsymbol{\beta}| \pm \frac{1}{2}\boldsymbol{R}_2^\bullet|\boldsymbol{\beta}|, p_2^\bullet; ...; \boldsymbol{C}_n^\bullet|\boldsymbol{\beta}| \pm \frac{1}{2}\boldsymbol{R}_n^\bullet|\boldsymbol{\beta}|, p_n^\bullet\}, \tag{3.28}$$

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

This can be expressed in a more condensed way by using the previously defined matrices $\boldsymbol{C}^\bullet$, $\boldsymbol{R}^\bullet$, and $\boldsymbol{p}_j^\bullet$, one of the columns of the matrix $\boldsymbol{P}^\bullet$. Considering that the resulting histogram is defined as the set of subintervals and probabilitites ($\boldsymbol{y}$, $\boldsymbol{p}$), we can represent the operation $\boldsymbol{\beta}^t\boldsymbol{x}$ as

$$\boldsymbol{\beta}^t\boldsymbol{x} = (\boldsymbol{C}^\bullet|\boldsymbol{\beta}| \pm \frac{1}{2}\boldsymbol{R}^\bullet|\boldsymbol{\beta}|, \boldsymbol{p}_j^\bullet), \tag{3.29}$$

or as the set of the centers, ranges, and probabilities ($\boldsymbol{c}$, $\boldsymbol{r}$, $\boldsymbol{p}$):

$$\boldsymbol{\beta}^t\boldsymbol{x} = (\boldsymbol{C}^\bullet|\boldsymbol{\beta}|, \boldsymbol{R}^\bullet|\boldsymbol{\beta}|, \boldsymbol{p}_j^\bullet). \tag{3.30}$$

These last expressions allow us to calculate any linear combination of histograms easily, as it can be seen in Example 3.2.7.

Example 3.2.7. *Considering the three histograms x_1, x_2, and x_3 described as follows:*

$$x_1 = \{[-2, 1[, 0.1; [1, 2[, 0.7; [2, 6], 0.2\},$$
$$x_2 = \{[2, 8[, 0.5; [8, 12[, 0.5\},\ x_3 = \{[0, 7[, 0.4; [7, 10[, 0.6\},$$

if we want to calculate a linear combination of these three histograms, for example,

$$x_1 + 3x_2 - 4x_3,$$

this linear combination is equivalent to the expression $\boldsymbol{\beta}^t\boldsymbol{x}$ with $\boldsymbol{\beta}^t = [1, 3, -4]$ and $\boldsymbol{x}^t = [x_1\, x_2\, x_3]$.

Firstly, a harmonization of these three histograms has to be performed. Noticing that $\beta_3 < 0$, it is necessary to first compute the inverse of x_3 before harmonizing it. Consequently, $-x_3$ is given by

$$-x_3 = \{[-10, -7[, 0.6; [-7, 0[, 0.4\}.$$

Now, when we do the harmonization of the histograms x_1, x_2, and the previously calculated $-x_3$, the result (rounded to two decimal places when needed) is:

$$x_1^\star = \{[-2, 1[, 0.1; [1, 1.57[, 0.4; [1.57, 1.71[, 0.1; [1.71, 2[, 0.2; [2, 6], 0.2\},$$
$$x_2^\star = \{[2, 3.2[, 0.1; [3.2, 8[, 0.4; [8, 8.8[, 0.1; [8.8, 10.4[, 0.2; [10.4, 12], 0.2\},$$
$$(-x_3)^\star = \{[-10, -9.5[, 0.1; [-9.5, -7.5[, 0.4; [-7.5, -7[, 0.1; [-7, -3.5[, 0.2; [-3.5, 0], 0.2\}.$$

Since $\beta_3 < 0$, it is necessary to be careful when computing the third column of the matrices $\boldsymbol{C}^\bullet$ and $\boldsymbol{R}^\bullet$. The centers and ranges of this column are taken from $(-x_3)^\star$. Thus, the matrices $\boldsymbol{C}^\bullet$, $\boldsymbol{R}^\bullet$, and $\boldsymbol{P}^\bullet$ for this example are:

$$\boldsymbol{C}^\bullet = \begin{bmatrix} -0.50 & 2.60 & -9.75 \\ 1.29 & 5.60 & -8.50 \\ 1.64 & 8.40 & -7.25 \\ 1.86 & 9.60 & -5.25 \\ 4.00 & 11.20 & -1.75 \end{bmatrix}, \boldsymbol{R}^\bullet = \begin{bmatrix} 3.00 & 1.20 & 0.50 \\ 0.57 & 4.80 & 2.00 \\ 0.14 & 0.80 & 0.50 \\ 0.29 & 1.60 & 3.50 \\ 4.00 & 1.60 & 3.50 \end{bmatrix}, \boldsymbol{P}^\bullet = \begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 \end{bmatrix}.$$

To obtain the histogram resulting from this operation, with the notation of the bounds of the subintervals, it is just necessary to use the expression in (3.29). By replacing the previously computed matrices $\boldsymbol{C}^\bullet$ and $\boldsymbol{R}^\bullet$ in the expression $\boldsymbol{C}^\bullet|\boldsymbol{\beta}| \pm \frac{1}{2}\boldsymbol{R}^\bullet|\boldsymbol{\beta}|$, we get the lower and upper bounds of the histogram subintervals:

$$\boldsymbol{\beta}^t\boldsymbol{x} = \begin{bmatrix} -0.50 & 2.60 & -9.75 \\ 1.29 & 5.60 & -8.50 \\ 1.64 & 8.40 & -7.25 \\ 1.86 & 9.60 & -5.25 \\ 4.00 & 11.20 & -1.75 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \pm \frac{1}{2} \begin{bmatrix} 3.00 & 1.20 & 0.50 \\ 0.57 & 4.80 & 2.00 \\ 0.14 & 0.80 & 0.50 \\ 0.29 & 1.60 & 3.50 \\ 4.00 & 1.60 & 3.50 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} [-36.00, -27.40[\\ [-27.40, -4.43[\\ [-4.43, 0.11[\\ [0.11, 19.20[\\ [19.20, 42.00] \end{bmatrix}.$$

Therefore, we have that the result of $\boldsymbol{\beta}^t\boldsymbol{x}$ in the form $(\boldsymbol{y}, \boldsymbol{p})$ is

$$\boldsymbol{\beta}^t\boldsymbol{x} = \left(\begin{bmatrix} [-36.00, -27.40[\\ [-27.40, -4.43[\\ [-4.43, 0.11[\\ [0.11, 19.20[\\ [19.20, 42.00] \end{bmatrix}, \begin{bmatrix} 0.1 \\ 0.4 \\ 0.1 \\ 0.2 \\ 0.2 \end{bmatrix} \right),$$

which is equivalent to the histogram x that is the solution of this linear combination problem:

$$x = \{[-36, -27.4[, 0.1; [-27.4, -4.43[, 0.4; [-4.43, 0.11[, 0.1; [0.11, 19.2[, 0.2; [19.2, 42], 0.2\}.$$

□

Having defined a condensed formula for the linear combination of histograms, it is also interesting to check some of its properties and whether or not this newly defined algebra for histogram variables can be considered as a vector space.

Before proceeding with the demonstrations of the axioms of a vector space, it is important to prove an important property of the harmonization procedure: the order by which the harmonization is performed

is irrelevant. Up to now, the harmonization was performed at once for all the histograms that participate in the operations. Hence, this property was inconsequential for the previous results. Nonetheless, in the next proofs, the order by which the histograms undergo this procedure is important. This requires the introduction of some new notation regarding the harmonization procedure. It is considered that, for a histogram represented as $x_j^{\star utz}$, the set of indices next to the $\star$ symbol represent which histograms were harmonized together with x_j and in which order. Therefore, for $x_1^{\star 324}$, the histogram x_1 was first harmonized with x_3; the result was subsequently harmonized with the histogram x_2 and, finally, a harmonization was performed on the result of all the previous operations with x_4. This same notation is used for the harmonized centers, ranges, and probabilities.

Proposition 3.2.1. *Considering x_1, x_2, and x_3 to be arbitrary histograms, the order by which these histograms are harmonized is irrelevant. That is,*

$$x_1^{\star 23} = x_1^{\star 32}.$$

Proof. Defining the histograms x_1, x_2, and x_3, such that:

$$\begin{aligned}
x_1 &= \{(c_{11}, r_{11}), p_{11}; ..; (c_{i1}, r_{i1}), p_{i1}; ..; (c_{n_1 1}, r_{n_1 1}), p_{n_1 1}\},\\
x_2 &= \{(c_{12}, r_{12}), p_{12}; ..; (c_{i2}, r_{i2}), p_{i2}; ..; (c_{n_2 2}, r_{n_2 2}), p_{n_2 2}\},\\
x_3 &= \{(c_{13}, r_{13}), p_{13}; ..; (c_{i3}, r_{i3}), p_{i3}; ..; (c_{n_3 3}, r_{n_3 3}), p_{n_3 3}\}.
\end{aligned}$$

The harmonization on the histograms x_1 and x_2 originates the following harmonized histogram $x_1^{\star 2}$ with n subintervals:

$$x_1^{\star 2} = (c_{11}^{\star 2}, r_{11}^{\star 2}), p_{11}^{\star 2}; ..; (c_{i1}^{\star 2}, r_{i1}^{\star 2}), p_{i1}^{\star 2}; ..; (c_{n1}^{\star 2}, r_{n1}^{\star 2}), p_{n1}^{\star 2}\}.$$

By harmonizing $x_1^{\star 2}$ with x_3, the histogram $x_1^{\star 23}$ with m subintervals ($m \geq n$) is created:

$$x_1^{\star 23} = (c_{11}^{\star 23}, r_{11}^{\star 23}), p_{11}^{\star 23}; ..; (c_{i1}^{\star 23}, r_{i1}^{\star 23}), p_{i1}^{\star 23}; ..; (c_{k1}^{\star 23}, r_{m}^{\star 23}), p_{m1}^{\star 23}\},$$

where $c_{i1}^{\star 23}$ and $r_{i1}^{\star 23}$ are given by

$$c_{i1}^{\star 23} = \frac{Q_{X_1}(w_{i-1\,1}^{\star 23}) + Q_{X_1}(w_{i1}^{\star 23})}{2},\ r_{i1}^{\star 23} = Q_{X_1}(w_{i-1\,1}^{\star 23}) - Q_{X_1}(w_{i1}^{\star 23}), \text{ with } i \in \{1, .., m\},$$

and where the $w_{i1}^{\star 23}$ are the cumulative probabilities computed through the expression

$$w_{i1}^{\star 23} = \begin{cases} 0, & i = 0 \\ \sum_{h=1}^{i} p_{h1}^{\star 23}, & i = 1, .., m \end{cases}.$$

Following the same procedure, in the case where the histogram x_1 is harmonized first with x_3 and the result is once again harmonized with x_2, the histogram $x_1^{\star 32}$ is produced:

$$x_1^{\star 32} = (c_{11}^{\star 32}, r_{11}^{\star 32}), p_{11}^{\star 32}; ..; (c_{i1}^{\star 32}, r_{i1}^{\star 32}), p_{i1}^{\star 32}; ..; (c_{n_1 1}^{\star 32}, r_{n_1 1}^{\star 32}), p_{n_1 1}^{\star 32}\},$$

whose the centers and ranges can be determined according to the expressions

$$c_{i1}^{\star 32} = \frac{Q_{X_1}(w_{i-1\,1}^{\star 32}) + Q_{X_1}(w_{i1}^{\star 32})}{2},\ r_{i1}^{\star 32} = Q_{X_1}(w_{i-1\,1}^{\star 32}) - Q_{X_1}(w_{i1}^{\star 32}), \text{ with } i \in \{1, .., k\},$$

with the cumulative probabilities $w_{i1}^{\star 32}$ defined by

$$w_{i1}^{\star 32}=\begin{cases}0, & i=0\\ \sum_{h=1}^{i} p_{h1}^{\star 32}, & i=1,..,k\end{cases}.$$

Comparing the expressions for both cases, it is possible to conclude that the condition $x_1^{\star 23} = x_1^{\star 32}$ depends solely on the set of the harmonized cumulative probabilities $\mathcal{W}^{\star 23}$ and $\mathcal{W}^{\star 32}$. If they are the same, all the centers, ranges and probabilities will also be equal in the expressions for $x_1^{\star 23}$ and $x_1^{\star 32}$. Since the sets $\mathcal{W}^{\star}$ are created by simply taking the non-repetitive cumulative probabilities w_i from the histograms that take part in the harmonization procedure, the order by which the process is performed does not affect the results and we have that $\mathcal{W}^{\star 23}$=$\mathcal{W}^{\star 32}$, and consequently, $x_1^{\star 23} = x_1^{\star 32}$. □

In the subsequent proofs, it is considered that $\mathcal{X}$ is the set of the histograms defined by their set of subintervals $\mathcal{Y}$ and the set of probability functions $\mathcal{P}$, such that:

$$\mathcal{X} = \mathcal{Y}(\mathbb{R}^2) \times \mathcal{P}, \text{ with } \mathcal{P} = \{\boldsymbol{p}_j : \boldsymbol{p}_j = (p_{ij},..,p_{n_j j})^t, \textstyle\sum_{i=1}^{n_j} p_{ij} = 1,\, 0 \leq p_{ij} \leq 1\}.$$

$\mathcal{X}$ can also be defined using the set of the centers and ranges of the subintervals, $\mathcal{C}$ and $\mathcal{R}$, together with the previously defined set of probabilities $\mathcal{P}$:

$$\mathcal{X} = \mathcal{C}(\mathbb{R}) \times \mathcal{R}(\mathbb{R}_0^+) \times \mathcal{P}.$$

The operations carried out over the set $\mathcal{X}$ follow the general expression in (3.30) and respect the properties presented in Tables 3.4 and 3.5. Now, it is necessary to check whether or not the operations defined by this expression satisfy the eight axioms of a vector space. To check their validity, the histogram x_j in the set $\mathcal{X}$ is represented, as in (2.8), by the vectors of the sets of the ordered centers ($\boldsymbol{c}_j$), ranges ($\boldsymbol{r}_j$) of its subintervals and the associated vector of the set of probabilities $\boldsymbol{p}_j$, such that for a histogram x_j: x_j=($\boldsymbol{c}_j$,$\boldsymbol{r}_j$,$\boldsymbol{p}_j$). The number of subintervals prior to the harmonization process may be different for each histogram.

Using the previous proposition and notations and, once again, considering x_1, x_2, and x_3 to be arbitrary histograms belonging to the set $\mathcal{X}$, we shall show that most, but not all, of the axioms of a vector space hold.

1. $x_1 + (x_2 + x_3) = (x_1 + x_2) + x_3$.

   Proof. The expression for x_1 + (x_2 + x_3) is given by

   $x_1+(x_2+x_3) = (\boldsymbol{c}_1,\boldsymbol{r}_1,\boldsymbol{p}_1)+((\boldsymbol{c}_2,\boldsymbol{r}_2,\boldsymbol{p}_2)+(\boldsymbol{c}_3,\boldsymbol{r}_3,\boldsymbol{p}_3)) = (\boldsymbol{c}_1,\boldsymbol{r}_1,\boldsymbol{p}_1)+((\boldsymbol{c}_2^{\star 3},\boldsymbol{r}_2^{\star 3},\boldsymbol{p}_2^{\star 3})+(\boldsymbol{c}_3^{\star 2},\boldsymbol{r}_3^{\star 2},\boldsymbol{p}_3^{\star 2})) = (\boldsymbol{c}_1,\boldsymbol{r}_1,\boldsymbol{p}_1) + (\boldsymbol{c}_2^{\star 3}+\boldsymbol{c}_3^{\star 2},\boldsymbol{r}_2^{\star 3}+\boldsymbol{r}_3^{\star 2},\boldsymbol{p}_2^{\star 3}) = (\boldsymbol{c}_1^{\star 23},\boldsymbol{r}_1^{\star 23},\boldsymbol{p}_1^{\star 23}) + (\boldsymbol{c}_2^{\star 31}+\boldsymbol{c}_3^{\star 21},\boldsymbol{r}_2^{\star 31}+\boldsymbol{r}_3^{\star 21},\boldsymbol{p}_2^{\star 31}) = (\boldsymbol{c}_1^{\star 23} + \boldsymbol{c}_2^{\star 31}+\boldsymbol{c}_3^{\star 21},\boldsymbol{r}_1^{\star 23}+\boldsymbol{r}_2^{\star 31}+\boldsymbol{r}_3^{\star 21},\boldsymbol{p}_1^{\star 23})$,

   where the fact that the probabilities resulting from a harmonization are the same for all the histograms involved ($\boldsymbol{p}_h^{\star k} = \boldsymbol{p}_k^{\star h}$) is used.

   Following the same reasoning, the expression for $(x_1 + x_2) + x_3$ is

   $(x_1+x_2)+x_3 = ((\boldsymbol{c}_1,\boldsymbol{r}_1,\boldsymbol{p}_1)+(\boldsymbol{c}_2,\boldsymbol{r}_2,\boldsymbol{p}_2))+(\boldsymbol{c}_3,\boldsymbol{r}_3,\boldsymbol{p}_3) = ((\boldsymbol{c}_1^{\star 2},\boldsymbol{r}_1^{\star 2},\boldsymbol{p}_1^{\star 2})+(\boldsymbol{c}_2^{\star 1},\boldsymbol{r}_2^{\star 1},\boldsymbol{p}_2^{\star 1}))+(\boldsymbol{c}_3,\boldsymbol{r}_3,\boldsymbol{p}_3) = (\boldsymbol{c}_1^{\star 2}+\boldsymbol{c}_2^{\star 1},\boldsymbol{r}_1^{\star 2}+\boldsymbol{r}_2^{\star 1},\boldsymbol{p}_1^{\star 2}) + (\boldsymbol{c}_3,\boldsymbol{r}_3,\boldsymbol{p}_3) = (\boldsymbol{c}_1^{\star 23}+\boldsymbol{c}_2^{\star 13},\boldsymbol{r}_1^{\star 23}+\boldsymbol{r}_2^{\star 13},\boldsymbol{p}_1^{\star 23}) + (\boldsymbol{c}_3^{\star 12},\boldsymbol{r}_3^{\star 12},\boldsymbol{p}_3^{\star 12}) = (\boldsymbol{c}_1^{\star 23} + \boldsymbol{c}_2^{\star 13}+\boldsymbol{c}_3^{\star 12},\boldsymbol{r}_1^{\star 23}+\boldsymbol{r}_2^{\star 13}+\boldsymbol{r}_3^{\star 12},\boldsymbol{p}_1^{\star 23})$.

Comparing both expressions and since, according to Proposition 3.2.1, $\boldsymbol{c}_2^{\star 31} = \boldsymbol{c}_2^{\star 13}$ and $\boldsymbol{r}_2^{\star 31} = \boldsymbol{r}_2^{\star 13}$, we have that $x_1 + (x_2 + x_3) = (x_1 + x_2) + x_3$. □

2. $x_1 + x_2 = x_2 + x_1$.

Proof. There are only two histograms involved in this condition; hence, it is not necessary to specify the order of the harmonization procedure. In addition, the regular notation $x^\star$ is used to represent the harmonized histogram x . As a result and recalling that $p_1^\star = p_2^\star$, the expressions for $x_1 + x_2$ and $x_2 + x_1$ are given by:

- $x_1 + x_2 = (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) + (\boldsymbol{c}_2, \boldsymbol{r}_2, \boldsymbol{p}_2) = (\boldsymbol{c}_1^\star, \boldsymbol{r}_1^\star, \boldsymbol{p}_1^\star) + (\boldsymbol{c}_2^\star, \boldsymbol{r}_2^\star, \boldsymbol{p}_2^\star) = (\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star, \boldsymbol{r}_1^\star + \boldsymbol{r}_2^\star, \boldsymbol{p}_1^\star)$
- $x_2 + x_1 = (\boldsymbol{c}_2, \boldsymbol{r}_2, \boldsymbol{p}_2) + (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) = (\boldsymbol{c}_2^\star, \boldsymbol{r}_2^\star, \boldsymbol{p}_2^\star) + (\boldsymbol{c}_1^\star, \boldsymbol{r}_1^\star, \boldsymbol{p}_1^\star) = (\boldsymbol{c}_2^\star + \boldsymbol{c}_1^\star, \boldsymbol{r}_2^\star + \boldsymbol{r}_1^\star, \boldsymbol{p}_1^\star)$

As both expressions are the same, $x_1 + x_2 = x_2 + x_1$. □

3. $\exists\, 0 \in \mathcal{X}$, such that $x_1 + 0 = x_1,\ \forall x_1 \in \mathcal{X}$.

Proof. Considering $0 = (0, 0, 1)$, which represents the single-value 0:

$x_1 + 0 = (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) + (0, 0, 1) = (\boldsymbol{c}_1 + 0, \boldsymbol{r}_1 + 0, \boldsymbol{p}_1) = (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) = x_1$.

Therefore, the proof is complete. □

4. For every $x_1 \in \mathcal{X}$, there exists an element $-x_1 \in \mathcal{X}$, such that $x_1 + (-x_1) = 0$. This axiom does not hold.

Proof. From the previous proof, it was defined that $0 = (0, 0, 1)$. In Example 3.2.6, it was observed that the operation $x_1 + (-x_1)$ does not equal 0; instead it results in a histogram symmetric with respect to the y-axis.

As for the arithmetic operations with quantile functions presented in Table 3.6 there never occurs a diminution of the value of the resulting ranges of the histograms, it is not possible to reach a range equal to 0, unless the initial value for the range is already 0. Thus, this condition is not met and this algebra does not specify a vector space. Nonetheless, it is still verified if this algebra satisfies the remaining axioms of a vector space. □

5. $a(bx_1) = (ab)x_1$, with $a, b \in \mathbb{R}$.

Proof. The expressions for $a(bx_1)$ and $(ab)x_1$ depend on the signs of the constants a and b. When a histogram is multiplied by a negative constant, the order of the subintervals is reversed. Hence, for the vectors of the centers, ranges and probabilities where this occurs, a notation with the symbol $\diamond$ is used. Thus,

$$\boldsymbol{c}_1^{\diamond}=\begin{bmatrix} c_{n_1 1} \\ c_{n_1-1\,1} \\ \vdots \\ c_{11} \end{bmatrix}, \boldsymbol{r}_1^{\diamond}=\begin{bmatrix} r_{n_1 1} \\ r_{n_1-1\,1} \\ \vdots \\ r_{11} \end{bmatrix}, \boldsymbol{p}_1^{\diamond}=\begin{bmatrix} p_{n_1 1} \\ p_{n_1-1\,1} \\ \vdots \\ p_{11} \end{bmatrix}.$$

The expression for bx_1 is given by

$$bx_1=\begin{cases} (b\boldsymbol{c}_1, b\boldsymbol{r}_1, \boldsymbol{p}_1), & b \geq 0 \\ (b\boldsymbol{c}_1^{\diamond}, |b|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & b < 0 \end{cases}.$$

Multiplying the previous expression by a, $a(bx_1)$ is obtained, such that:

$$a(bx_1)=\begin{cases} (ab\boldsymbol{c}_1, ab\boldsymbol{r}_1, \boldsymbol{p}_1), & a \geq 0, b \geq 0 \\ (ab\boldsymbol{c}_1^{\diamond}, a|b|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a \geq 0, b < 0 \\ (ab\boldsymbol{c}_1^{\diamond}, |a|b\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a < 0, b \geq 0 \\ (ab\boldsymbol{c}_1, |ab|\boldsymbol{r}_1, \boldsymbol{p}_1), & a < 0, b < 0 \end{cases}.$$

The operation $(ab)x_1$ depends on whether the product ab is positive or negative. Hence,

$$(ab)x_1=\begin{cases} (ab\boldsymbol{c}_1, ab\boldsymbol{r}_1, \boldsymbol{p}_1), & a \geq 0, b \geq 0 \\ (ab\boldsymbol{c}_1^{\diamond}, |ab|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a \geq 0, b < 0 \\ (ab\boldsymbol{c}_1^{\diamond}, |ab|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a < 0, b \geq 0 \\ (ab\boldsymbol{c}_1, ab\boldsymbol{r}_1, \boldsymbol{p}_1), & a < 0, b < 0 \end{cases}.$$

By analyzing the signs of the constants in both expressions, it can be concluded that they are equal. Hence, $a(bx_1) = (ab)x_1$ with $a, b \in \mathbb{R}$. □

6. $1x_1 = x_1$, $\forall x_1 \in \mathcal{X}$.

Proof. By computing the product of 1 and a histogram x_1, we get

$$1 \times x_1 = 1 \times (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) = (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1).$$

Hence, the proof is complete. □

7. $a(x_1 + x_2) = ax_1 + ax_2$, with $a \in \mathbb{R}$.

Proof. Once again, the notation with the symbol $\diamond$ is used, representing the vectors where the order of the subintervals is reversed. But, in this case, it is important to differentiate the order by which the reversion of the subintervals and the harmonization occurs. Therefore, $\boldsymbol{c}^{\diamond\star}$ corresponds to the set of centers which first had their order reversed and were subsequently harmonized, while $\boldsymbol{c}^{\star\diamond}$ is the representation when a harmonization is performed before the order of the subintervals is reversed. The same logic is applied to the sets of ranges and probabilities.

The expression for $x_1 + x_2$ is given by

$$x_1 + x_2 = (\boldsymbol{c}_1, \boldsymbol{r}_1, \boldsymbol{p}_1) + (\boldsymbol{c}_2, \boldsymbol{r}_2, \boldsymbol{p}_2) = (\boldsymbol{c}_1^\star, \boldsymbol{r}_1^\star, \boldsymbol{p}_1^\star) + (\boldsymbol{c}_2^\star, \boldsymbol{r}_2, \boldsymbol{p}_2^\star) = (\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star, \boldsymbol{r}_1^\star + \boldsymbol{r}_2, \boldsymbol{p}_1^\star).$$

The product of this histogram and the constant a is

$$a(x_1 + x_2) = \begin{cases} (a(\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star), a(\boldsymbol{r}_1^\star + \boldsymbol{r}_2^\star), \boldsymbol{p}_1^\star), & a \geq 0 \\ (a(\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star)^\diamond, |a|(\boldsymbol{r}_1^\star + \boldsymbol{r}_2^\star)^\diamond, \boldsymbol{p}_1^{\star\diamond}), & a < 0 \end{cases}.$$

On the other hand, when the products ax_1 and ax_2 are computed before the addition is done, we have

$$ax_1 = \begin{cases} (a\boldsymbol{c}_1, a\boldsymbol{r}_1, \boldsymbol{p}_1), & a \geq 0 \\ (a\boldsymbol{c}_1^\diamond, |a|\boldsymbol{r}_1^\diamond, \boldsymbol{p}_1^\diamond), & a < 0 \end{cases}, ax_2 = \begin{cases} (a\boldsymbol{c}_2, a\boldsymbol{r}_2, \boldsymbol{p}_2), & a \geq 0 \\ (a\boldsymbol{c}_2^\diamond, |a|\boldsymbol{r}_2^\diamond, \boldsymbol{p}_2^\diamond), & a < 0 \end{cases}.$$

By performing a harmonization procedure on both of the previous histograms, we get

$$(ax_1)^\star = \begin{cases} (a\boldsymbol{c}_1^\star, a\boldsymbol{r}_1^\star, \boldsymbol{p}_1^\star), & a \geq 0 \\ (a\boldsymbol{c}_1^{\diamond\star}, |a|\boldsymbol{r}_1^{\diamond\star}, \boldsymbol{p}_1^{\diamond\star}), & a < 0 \end{cases}, (ax_2)^\star = \begin{cases} (a\boldsymbol{c}_2^\star, a\boldsymbol{r}_2^\star, \boldsymbol{p}_2^\star), & a \geq 0 \\ (a\boldsymbol{c}_2^{\diamond\star}, |a|\boldsymbol{r}_2^{\diamond\star}, \boldsymbol{p}_2^{\diamond\star}), & a < 0 \end{cases},$$

and then, by adding both of the histograms, the expression for $ax_1 + ax_2$ is obtained:

$$ax_1 + ax_2 = \begin{cases} (a(\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star), a(\boldsymbol{r}_1^\star + \boldsymbol{r}_2^\star), \boldsymbol{p}_1^\star), & a \geq 0 \\ (a(\boldsymbol{c}_1^{\diamond\star} + \boldsymbol{c}_2^{\diamond\star}), |a|(\boldsymbol{r}_1^{\diamond\star} + \boldsymbol{r}_2^{\diamond\star}), \boldsymbol{p}_1^{\diamond\star}), & a < 0 \end{cases}.$$

The comparison of the expressions for $a(x_1+x_2)$ and ax_1+ax_2, allows us to infer that it is necessary to check if $(\boldsymbol{c}_1^\star + \boldsymbol{c}_2^\star)^\diamond = \boldsymbol{c}_1^{\diamond\star} + \boldsymbol{c}_2^{\diamond\star}$, $(\boldsymbol{r}_1^\star + \boldsymbol{r}_2^\star)^\diamond = \boldsymbol{r}_1^{\diamond\star} + \boldsymbol{r}_2^{\diamond\star}$ and $\boldsymbol{p}_1^{\diamond\star} = \boldsymbol{p}_1^{\star\diamond}$.

It was previously observed that the values of the centers and ranges depend solely on the cumulative probabilities. Hence, we only need to check the condition $\mathcal{W}^{\diamond\star} = \mathcal{W}^{\star\diamond}$. When harmonizing the vectors $\boldsymbol{p}_1^\diamond = [p_{n_1 1}, .., p_{11}]^t, \boldsymbol{p}_2^\diamond = [p_{n_2 2}, .., p_{22}]^t$, the cumulative probability set $\mathcal{W}^{\diamond\star}$ is created by taking the non-repeated cumulative probabilities from the sets $\mathcal{W}_1^\diamond$ and $\mathcal{W}_2^\diamond$. The elements of these two sets $w_{i1}^\diamond$ and $w_{i2}^\diamond$ are, respectively:

$$w_{i1}^\diamond = \begin{cases} 0, & i = 0 \\ \sum_{h=1}^{i} p_{h1}^\diamond, & i = 1, .., k \end{cases}, w_{i2}^\diamond = \begin{cases} 0, & i = 0 \\ \sum_{h=1}^{i} p_{h2}^\diamond, & i = 1, .., k \end{cases}.$$

If the set $\mathcal{W}^\star$ is created first by taking the non-repeated elements from the original sets $\mathcal{W}_1$ and $\mathcal{W}_2$, and then their order is reversed, it will also produce the same elements as for $\mathcal{W}^{\diamond\star}$. Therefore, $\mathcal{W}^{\diamond\star} = \mathcal{W}^{\star\diamond}$ and $a(x_1 + x_2) = ax_1 + ax_2$, with $a \in \mathbb{R}$. □

8. $(a + b)x_1 = ax_1 + bx_1$, with $a, b \in \mathbb{R}$. This axiom does not hold.

Proof. The expression for $(a + b)x_1$ is given by

$$(a+b)x_1 = \begin{cases} ((a+b)\boldsymbol{c}_1, (a+b)\boldsymbol{r}_1, \boldsymbol{p}_1), & a + b \geq 0 \\ ((a+b)\boldsymbol{c}_1, |a+b|\boldsymbol{r}_1, \boldsymbol{p}_1), & a + b < 0 \end{cases}.$$

On the other hand, ax_1 and bx_1 are given by

$$ax_1 = \begin{cases} (a\boldsymbol{c}_1, a\boldsymbol{r}_1, \boldsymbol{p}_1), & a \geq 0 \\ (a\boldsymbol{c}_1^{\diamond}, |a|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a < 0 \end{cases}, \; bx_1 = \begin{cases} (b\boldsymbol{c}_1, b\boldsymbol{r}_1, \boldsymbol{p}_1), & b \geq 0 \\ (b\boldsymbol{c}_1^{\diamond}, |b|\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & b < 0 \end{cases}.$$

The result of the operation $ax_1 + bx_1$ is

$$ax_1 + bx_1 = \begin{cases} ((a+b)\boldsymbol{c}_1, (a+b)\boldsymbol{r}_1, \boldsymbol{p}_1), & a \geq 0, b \geq 0 \\ (a\boldsymbol{c}_1^{\diamond\star} + b\boldsymbol{c}_1^{\star}, |a|\boldsymbol{r}_1^{\diamond\star} + b\boldsymbol{r}_1^{\star}, \boldsymbol{p}_1^{\diamond\star}), & a < 0, b \geq 0 \\ (a\boldsymbol{c}_1^{\star} + b\boldsymbol{c}_1^{\diamond\star}, a\boldsymbol{r}_1^{\star} + |b|\boldsymbol{r}_1^{\diamond\star}, \boldsymbol{p}_1^{\diamond\star}), & a \geq 0, b < 0 \\ ((a+b)\boldsymbol{c}_1^{\diamond}, (|a|+|b|)\boldsymbol{r}_1^{\diamond}, \boldsymbol{p}_1^{\diamond}), & a < 0, b < 0 \end{cases}.$$

The expressions for $(a+b)x_1$ and $ax_1 + bx_1$ are different, so this axiom is also not valid for this algebra. □

Consequently, it is possible to conclude that this algebra for histograms does not represent a vector space. It fails the fourth and eighth axioms previously stated.

It is also interesting to check how this algebra behaves in the particular case of the interval variables, as well as its relation with Moore's Interval Algebra. An interval is the particular case of a histogram with only one subinterval, which has an associated probability of 1. Thus, using the notation of histograms for an interval z, we have that

$$z = \{y, p\}, \text{ with } y = [a, b],\ a,b \in \mathbb{R},\ a \leq b, \text{ and } p = 1.$$

It is not necessary to perform the harmonization procedure with interval variables, since the same probability is associated to all of them, which is equal to one. So, the notations with the symbols $\star$, $\diamond$, and $\bullet$ are also useless, and all the operations involving interval variables using this algebra follow the same rules as the ones presented in Table 3.1 for Moore's Interval Algebra.

By applying the expression (3.29) for the interval case, we get

$$\boldsymbol{\beta}^t \boldsymbol{z} = (\boldsymbol{c}\boldsymbol{\beta} \pm \tfrac{1}{2}\boldsymbol{r}|\boldsymbol{\beta}|, p_j), \text{ with } p_j = 1.$$

When comparing this expression to Moore's Algebra, the only difference is that, under the notation of the previously defined histogram algebra, the linear combinations of k intervals are now done with the row-vectors $\boldsymbol{c}$, $\boldsymbol{r}$ with dimensions $(1 \times k)$, instead of the column-vectors defined in Moore's Interval Algebra. However, the results for intervals are still equal. Hence, it is possible to conclude that the histogram algebra with quantile functions is the same as Moore's Interval Algebra, when only interval variables are involved in the operations. Hence, Moore's Interval Algebra can be seen as a special case of the histogram algebra proposed here.

Chapter 4

Descriptive Statistics

The last goal of this work is to apply the previously defined algebra for histograms with quantile functions to SPCA. However, before proceeding with this task, it is important to present the definitions of some statistical measures for both interval and histogram variables: sample symbolic mean, variance, covariance, and correlation. Among these measures, the most relevant one for this work is the sample symbolic covariance. Some of the definitions of sample symbolic covariance presented in this chapter are used in Chapter 5 to build sample symbolic covariance matrices for the SPCA with histogram-valued variables.

4.1 Sample symbolic mean and variance

Before starting to define any statistical measure, it is necessary to characterize the empirical distribution function for histogram variables, $F(\xi)$, and the corresponding density function $f(\xi)$, from which some of the statistical measures that will be presented next have been deduced. These functions were introduced in Chapter 6 of [4].

Considering that we have a set of k histograms x_j, each one characterized by n_j sequential subintervals $y_{ij} = [a_{ij}, b_{ij}] = (c_{ij}, r_{ij})$, where the micro-data within a given subinterval follows a uniform distribution, and with associated probabilities p_{ij}, the empirical distribution function $F(\xi)$ associated to this set is given by the expression

$$F(\xi) = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} P(A \leq \xi | A \in y_{ij}) p_{ij}, \ \xi \in \mathbb{R}, \tag{4.1}$$

with

$$P(y_{ij} \leq \xi) = \begin{cases} 0, & \xi < a_{ij} \\ \frac{\xi - a_{ij}}{b_{ij} - a_{ij}}, & b_{ij} \leq \xi < a_{ij} \\ 1, & \xi \geq b_{ij} \end{cases}.$$

By computing the derivative of the expression (4.1), we obtain the following formula for the corre-

sponding density function, $f(\xi)$:

$$f(\xi) = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{\mathbf{1}_{y_{ij}(\xi)}}{r_{ij}}p_{ij},\ \xi \in \mathbb{R}, \tag{4.2}$$

with $\mathbf{1}_{y_{ij}}(\xi)$=$\begin{cases}0, & \xi \notin y_{ij}\\ 1, & \xi \in y_{ij}\end{cases}$.

Now that $f(\xi)$ has been defined as the empirical distribution function over a set of k histogram variables, the next step is to present the definitions of some statistical measures. To do this, it is considered that we are dealing with a sample of k histograms x_j belonging to a histogram variable X, $(x_1, x_2, .., x_j, ..., x_k)$, whose associated micro-data within the subintervals y_{ij} follows a uniform distribution. The subintervals y_{ij} can be represented through their upper and lower bounds ($y_{ij} = [a_{ij}, b_{ij}]$), or by using their respective center and range ($y_i = (c_{ij}, r_{ij})$). It is important to remark that the number of subintervals in each histogram x_j may vary: the number of subintervals of the histogram x_j is considered to be equal to n_j.

In the subsequent definitions, the results for a histogram variable are presented first, and then the particular case of the interval variable is considered, which occurs when $n_j = 1, \forall j$, with $j = \{1, ..., k\}$ and associated probabilities $p_{1j} = 1$.

Under these conditions, the general formula for the sample symbolic mean of a histogram variable X based on k observations, $\bar{x}$, can be defined according to [12]. The expression of $\bar{x}$ can be deduced from (4.2), such that: $\bar{x} = \int_{-\infty}^{\infty}\xi f(\xi)d\xi$. By developing this formula, the following expression for $\bar{x}$, in terms of the bounds of the subintervals, is obtained:

$$\bar{x} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{a_{ij}+b_{ij}}{2}p_{ij}. \tag{4.3}$$

It can easily be perceived from the previous expression that this definition of sample symbolic mean corresponds to a regular weighted mean applied to the centers of the subintervals of the histograms. Thus, expression (4.3) can also be represented using the centers of the intervals, c_{ij}, such that:

$$\bar{x} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}c_{ij}p_{ij}. \tag{4.4}$$

To obtain the sample symbolic mean of an interval-valued variable Y based on a sample of size k, $\bar{y}$, it is just necessary to use $n_j = 1$ and $p_{ij} = 1$ in (4.4), which results in the expression

$$\bar{y} = \frac{1}{k}\sum_{j=1}^{k}c_j = \bar{c}. \tag{4.5}$$

This is generally the only definition of sample symbolic mean that is used, when the assumption of uniformly distributed histogram and interval-valued variables is adopted.

However, when it comes to the sample symbolic variance, several different definitions are used. The

first definition of sample symbolic variance presented here, which was introduced in [12], is denoted by s_1^2, having been deduced from $f(\xi)$ in (4.2), such that $s_1^2 = \int_{-\infty}^{\infty}(\xi - \bar{x})^2 f(\xi)d\xi$. By developing this expression, the following expression for s_1^2 is obtained:

$$s_1^2 = \frac{1}{3k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(b_{ij}^2 + b_{ij}a_{ij} + a_{ij}^2)p_{ij} - \bar{x}^2. \tag{4.6}$$

This expression can also be represented by using the notation of the centers and ranges:

$$s_1^2 = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}c_{ij}^2 p_{ij} - \bar{x}^2 + \frac{1}{12k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}r_{ij}^2 p_{ij}. \tag{4.7}$$

Unlike what happened for the sample symbolic mean, this expression does not correspond to the conventional definition of variance when applied to the weighted centers of the subintervals.

In the case of an interval variable Y, s_1^2 is given by

$$s_1^2 = \frac{1}{k}\sum_{j=1}^{k}c_j^2 - \bar{y}^2 + \frac{1}{12k}\sum_{j=1}^{k}r_j^2 = \frac{1}{k}\sum_{j=1}^{k}(c_j - \bar{c})^2 + \frac{1}{12}\sum_{j=1}^{k}\frac{r_j^2}{k}, \tag{4.8}$$

which corresponds to the sample symbolic variance of interval data $s_{jj}^{(\frac{1}{12})}$, according to [13].

The next definition of sample symbolic variance of histogram variables was proposed in [14] and is denoted by s_2^2. This definition corresponds to the conventional definition of variance, when applied to the mean of the weighted centers of each histogram x_j, $\bar{x}_j : \bar{x}_j = \sum_{i=1}^{n_j} c_{ij}p_{ij}$. Therefore, its expression is the following:

$$s_2^2 = \frac{1}{k}\sum_{j=1}^{k}(\sum_{i=1}^{n_j}c_{ij}p_{ij})^2 - \bar{x}^2 = \frac{1}{k}\sum_{j=1}^{k}\bar{x}_j^2 - \bar{x}^2. \tag{4.9}$$

For an interval variable Y, the expression for s_2^2 is

$$s_2^2 = \frac{1}{k}\sum_{j=1}^{k}c_j^2 - \bar{y}^2 = \frac{1}{k}\sum_{j=1}^{k}(c_j - \bar{c})^2, \tag{4.10}$$

which, according to [13], corresponds to the sample symbolic variance of interval data $s_j^{(0)}$, and ignores the contribution of the ranges.

Another definition of sample symbolic variance for interval variables, introduced in [15], measures the variability of the bounds of the intervals with respect to the sample symbolic mean. This sample symbolic variance is denoted by s_3^2. However, s_3^2 is still not defined for the case of the histogram variables and is only valid for interval variables. Thus, for an interval variable Y, s_3^2 is given by

$$s_3^2 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{(a_j - \bar{y})^2}{2} + \frac{(b_j - \bar{y})^2}{2}\right]. \tag{4.11}$$

The equivalent of the previous expression with the centers and ranges is given by

$$s_3^2 = \frac{1}{k}\sum_{i=1}^{k}(c_j - \bar{y})^2 + \frac{1}{4k}\sum_{j=1}^{k} r_j^2. \tag{4.12}$$

In [16], the relationship between these three definitions of sample symbolic variances was summarized for the interval-valued variables. Accordingly, if the expressions of s_1^2, s_2^2, and s_3^2 are compared for the interval case, it is possible to infer that s_1^2 and s_3^2 can be expressed in terms of s_2^2, which corresponds to the sample symbolic variance of the centers, and another term related to the expected value of the ranges of the intervals. As a consequence of this, we have

$$s_1^2 = s_2^2 + \tfrac{1}{12k}\textstyle\sum_{j=1}^{k} r_j^2,$$

$$s_3^2 = s_2^2 + \tfrac{1}{4k}\textstyle\sum_{j=1}^{k} r_j^2.$$

Considering that we have the random vectors related to the centers, **C**, and ranges, **R**, that $Var(\mathbf{C})$ corresponds to the variance of the centers of the intervals, and $E(\mathbf{R})$ to the expected value of the ranges, by applying the weak law of large numbers to these last results, we obtain

$$\begin{gathered} s_2^2 \xrightarrow{a.s} Var(\mathbf{C}), \\ s_1^2 \xrightarrow{a.s} Var(\mathbf{C}) + \tfrac{1}{12}E(\mathbf{R}^2), \\ s_3^2 \xrightarrow{a.s} Var(\mathbf{C}) + \tfrac{1}{4}E(\mathbf{R}^2). \end{gathered}$$

Nonetheless, we could not find the same kind of relationship between the definitions of sample symbolic variance s_1^2 and s_2^2.

For the previous definitions ($\bar{x}$, s_1^2, s_2^2), which are defined for histogram variables, it is also useful to describe their expressions in matrix form. By doing so, it becomes easier to compute these measures on real data. To achieve this, the matrices of the centers, ranges and probabilities associated to the subintervals from the k observations of a histogram variable X, represented by $\boldsymbol{C}$, $\boldsymbol{R}$, and $\boldsymbol{P}$, respectively, are considered:

$$\boldsymbol{C}=\begin{bmatrix} c_{11} & \dots & c_{1j} & \dots & c_{1k} \\ c_{21} & \dots & c_{2j} & \dots & c_{2k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{i1} & \dots & c_{ij} & \dots & c_{ik} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{n_1 1} & \dots & c_{n_1 j} & \dots & c_{n_1 k} \\ \vdots & \dots & \vdots & \dots & \vdots \end{bmatrix}, \boldsymbol{R}=\begin{bmatrix} r_{11} & \dots & r_{1j} & \dots & r_{1k} \\ r_{21} & \dots & r_{2j} & \dots & r_{2k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{i1} & \dots & r_{ij} & \dots & r_{ik} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{n_1 1} & \dots & r_{n_1 j} & \dots & r_{n_1 k} \\ \vdots & \dots & \vdots & \dots & \vdots \end{bmatrix}, \boldsymbol{P}=\begin{bmatrix} p_{11} & \dots & p_{1j} & \dots & p_{1k} \\ p_{21} & \dots & p_{2j} & \dots & p_{2k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \dots & p_{ij} & \dots & p_{ik} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n_1 1} & \dots & p_{n_1 j} & \dots & p_{n_1 k} \\ \vdots & \dots & \vdots & \dots & \vdots \end{bmatrix}.$$

As the number of subintervals of the k histograms can vary, these matrices need some adjustments. Their size is $(max(n_j) \times k)$, with $j = \{1, .., k\}$. For the matrix $\boldsymbol{P}$, the elements p_{ij} where $i > n_j$ have value 0, so that a probability zero is associated with any element of the matrices $\boldsymbol{C}$ and $\boldsymbol{R}$ that is not related to an actual subinterval of the histogram x_j. Since a probability of zero is associated with the elements c_{ij} and r_{ij} where $i > n_j$, their value is not relevant and can be chosen arbitrarily.

Using these previously defined matrices, it is possible to represent s_1^2 and s_2^2, for the histogram variable X, in the following way:

$$s_1^2 = \frac{1}{k}(tr[(\boldsymbol{C}^{(2)})^t \boldsymbol{P}]) - \frac{1}{k^2}[tr(\boldsymbol{C}^t \boldsymbol{P})]^2 + \frac{1}{12k}(tr[(\boldsymbol{R}^{(2)})^t \boldsymbol{P}], \tag{4.13}$$

$$s_2^2 = \frac{1}{k}(tr[(\boldsymbol{C}^t \boldsymbol{P})^{(2)}] - \frac{1}{k^2}[tr(\boldsymbol{C}^t \boldsymbol{P})]^2, \tag{4.14}$$

where $tr(\boldsymbol{B})$, represents the trace of the matrix $\boldsymbol{B}$, and $\boldsymbol{B}^{(2)}$ represents a matrix where the (i, j)-th entrance is b_{ij}^2 (and not the product $\boldsymbol{B} \times \boldsymbol{B}$).

4.2 Sample symbolic covariance and correlation

Next, three different definitions of sample symbolic covariance are presented. The first definition, designated as cov_1, was introduced in [4] and it was derived from the empirical joint density function of a sample of size k of two histogram variables X_1 and X_2, $(x_{11}, .., x_{j1}, ..., x_{k1})$ and $(x_{12}, x_{j2}, ..., x_{k2})$, respectively. The observation j of the histogram variable X_l, x_{jl}, with $l \in \{1, 2\}$, can be represented, under the notation of the bounds of its subintervals, by

$$x_{jl} = \{[a_{1jl}, b_{1jl}], p_{1jl}; ..; [a_{ijl}, b_{ijl}], p_{ijl}; ..; [a_{n_{jl}jl}, b_{n_{jl}jl}], p_{n_{jl}jl}\}, \text{ with } i \in \{1, .., n_{jl}\},\ j \in \{1, .., k\},$$

and, with the notation where the subintervals are sets of their centers and ranges,

$$x_{jl} = \{(c_{1jl}, r_{1jl}), p_{1jl}; ..; (c_{ijl}, r_{ijl}), p_{ijl}; ..; (c_{n_{jl}jl}, r_{n_{jl}jl}), p_{n_{jl}jl}\}, \text{ with } i \in \{1, .., n_{jl}\},\ j \in \{1, .., k\}.$$

Each histogram x_{j1} and x_{j2} can have a different number of subintervals. This definition of sample symbolic covariance, cov_1, corresponds to the case where the conventional expression of covariance is applied to the weighted means of each histogram j from the sample of the variables X_1 and X_2 (for the histogram j in $X_1 : \bar{x}_{j1} = \sum_{i_1=1}^{n_{j1}} c_{i_1 j1} p_{i_1 j1}$, and for the histogram j in $X_2 : \bar{x}_{j2} = \sum_{i_2=1}^{n_{j2}} c_{i_2 j2} p_{i_2 j2}$). Thus, the expression of $cov_1(X_1, X_2)$, for the histogram variables X_1 and X_2, is given by

$$cov_1(X_1, X_2) = \frac{1}{k} \sum_{j=1}^{k} \sum_{i_1=1}^{n_{j1}} c_{i_1 j1} p_{i_1 j1} \sum_{i_2=1}^{n_{j2}} c_{i_2 j2} p_{i_2 j2} - \bar{x}_1 \bar{x}_2 = \frac{1}{k} \sum_{j=1}^{k} \bar{x}_{j1} \bar{x}_{j2} - \bar{x}_1 \bar{x}_2. \tag{4.15}$$

For the interval case, the expression is given by

$$cov_1(Y_1, Y_2) = \frac{1}{k} \sum_{j=1}^{k} c_{j1} c_{j2} - \bar{y}_1 \bar{y}_2. \tag{4.16}$$

If $cov_1(X_1, X_2)$ is computed for two equal histogram variables ($X_1 = X_2$), the same expression as in (4.9) for s_2^2 is obtained. Hence, $s_2^2 = cov_1(X, X)$.

The second definition of sample symbolic covariance presented in this work, cov_2, was proposed in [17] and it was deduced by using an analogy that involves the symbolic variance s_1^2 and the similarity between the conventional expressions of variance and covariance. Nonetheless, $cov_2(X, X) \neq s_1^2$. The

expression of $cov_2(X_1, X_2)$ is more complex than usual, and for histogram variables is given by

$$cov_2(X_1, X_2) = \frac{1}{3k} \sum_{j=1}^{k} \sum_{i_1=1}^{n_{j1}} \sum_{i_2=1}^{n_{j2}} p_{i_1 j1} p_{i_2 j2} G_{i_1 j1} G_{i_2 j2} [Q_{i_1 j1} Q_{i_2 j2}]^{\frac{1}{2}}, \tag{4.17}$$

with

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

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

where $l \in \{1, 2\}$.

The introduction of the terms $G_{i_l jl}$ in the previous expression was done to guarantee that $Cov_2(X_1, X_2)$ is always a positive quantity. The terms $Q_{i_l jl}$ can also be represented in notation with the centers and ranges:

$$Q_{i_l jl} = 3c_{i_l jl}^2 + 3\bar{x}_l^2 - 6c_{i_l jl}\bar{x}_l + (\tfrac{r_{i_l jl}}{2})^2.$$

For two interval variables Y_1 and Y_2, the expression of this covariance is

$$cov_2(Y_1, Y_2) = \frac{1}{3k} \sum_{j=1}^{k} G_{j1} G_{j2} [Q_{j1} Q_{j2}]^{\frac{1}{2}}, \tag{4.18}$$

with

$$Q_{jl} = (a_{jl} - \bar{y}_l)^2 + (a_{jl} - \bar{y}_l)(b_{jl} - \bar{y}_l) + (b_{jl} - \bar{y}_l)^2,$$

$$G_{jl} = \begin{cases} -1, & c_{jl} \leq \bar{y}_l \\ 1, & c_{jl} > \bar{y}_l \end{cases}.$$

The third and last definition of sample symbolic covariance presented in this work, cov_3, was introduced in [18] and is only valid for interval variables. Its deduction was based on the decomposition of the Total Sum of Products (TSP) of two interval variables, Y_1 and Y_2, into the respective Within Sum of Products (WSP) and Between Sum of Products (BSP), such that:

$$TSP = WSP + BSP = k \times cov_3(Y_1, Y_2),$$

and where the WSP and BSP of two interval variables, Y_1 and Y_2, are defined in the following way:

$$WSP = \tfrac{1}{12} \textstyle\sum_{j=1}^{k} (b_{j1} - a_{j1})(b_{j2} - a_{j2}),$$

$$BSP = \textstyle\sum_{j=1}^{k} (c_{j1} - \bar{y}_1)(c_{j2} - \bar{y}_2)\ .$$

Hence, we have that the expression of $cov_3(Y_1, Y_2)$ is given by

$$\begin{aligned} cov_3(Y_1, Y_2) = \frac{1}{6k} \sum_{j=1}^{k} [&2(a_{j1} - \bar{y}_1)(a_{j2} - \bar{y}_2) + (a_{j1} - \bar{y}_1)(b_{j2} - \bar{y}_2) + \\ &+ (b_{j1} - \bar{y}_1)(a_{j2} - \bar{y}_2) + 2(b_{j1} - \bar{y}_1)(a_{j2} - \bar{y}_2). \end{aligned} \tag{4.19}$$

When the center and range notation is used, we get

$$cov_3(Y_1, Y_2) = \frac{1}{k}\sum_{j=1}^{k} c_{j1}c_{j2} - \bar{y}_1\bar{y}_2 + \frac{1}{k}\sum_{j=1}^{n}\frac{r_{j1}r_{j2}}{12}. \tag{4.20}$$

It is also important to mention that, for the particular case $Y_1 = Y_2$, we have that $cov_3(Y, Y) = s_3^2$.

Comparing the expression of cov_3 in (4.20) with the expression of cov_1 presented in (4.16) for interval variables, it can be ascertained that it is possible to rewrite $cov_3(Y_1, Y_2)$ in terms of cov_1:

$$cov_3(Y_1, Y_2) = cov_1(Y_1, Y_2) + \frac{1}{k}\sum_{j=1}^{n}\frac{r_{j1}r_{j2}}{12}. \tag{4.21}$$

It can be concluded that, for interval variables, all these definitions of sample symbolic variances and covariances (with the exception of cov_2) can be written in function of each other, whereas we were unable to find such a relationship for the histogram variables.

It is also useful to represent the expression of cov_1 (4.15) in matrix form, for histogram variables. The same is not done for cov_2 because its expression is too complex. By using the same definition of the matrices $\boldsymbol{C}$, $\boldsymbol{R}$, and $\boldsymbol{P}$, which were used for the symbolic variances, and by applying them to two histogram variables, X_1 and X_2, the matrices $\boldsymbol{C}_1$, $\boldsymbol{R}_1$, $\boldsymbol{P}_1$ and $\boldsymbol{C}_2$, $\boldsymbol{R}_2$, $\boldsymbol{P}_2$ are created and the following expression is obtained:

$$cov_1(X_1, X_2) = \frac{1}{k}(diag(\boldsymbol{C}_1^t\boldsymbol{P}_1))^t(diag(\boldsymbol{C}_2^t\boldsymbol{P}_2)) - \frac{1}{k^2}tr(\boldsymbol{C}_1^t\boldsymbol{P}_1)tr(\boldsymbol{C}_2^t\boldsymbol{P}_2), \tag{4.22}$$

where $diag(\boldsymbol{B})$ denotes a vector corresponding to the main diagonal of the matrix $\boldsymbol{B}$.

Another statistical measure that is important to define is the sample symbolic correlation between two histogram variables X_1 and X_2. The sample symbolic correlation, $r(X_1, X_2)$, is given by the general expression

$$r(X_1, X_2) = \frac{cov(X_1, X_2)}{\sqrt{s_{X_1}^2 s_{X_2}^2}}. \tag{4.23}$$

This expression for the correlation is valid for any definition of sample symbolic covariance and variance previously presented. However, it is obvious that it only makes sense to compute the sample correlation with sample covariances and variances that are related to each other. The relationships between the sample symbolic covariances and variances previously defined are:

- cov_1 (4.15) is related with s_1^2 (4.7), since both were obtained from density functions; it is also related directly with s_2^2 (4.9), since $cov_1(X, X) = s_2^2$.

- cov_2 (4.18) is related with s_1^2 (4.7), since its expression was derived from an analogy using this definition of symbolic variance.

- $cov_3(Y, Y) = s_3^2$ ((4.20) and (4.12)), but both of these measures are not yet generalized for histogram variables.

4.3 Example

An example is now presented, where the above mentioned definitions of the symbolic statistical measures are computed for a real data set.

The data set chosen for this task was taken from [19] and it 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 related to the their size: sepal width, petal width, petal length, and sepal length. By applying a K-means on this conventional data, according to [19], 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. By doing this procedure, the following histogram-valued variables were created:

- X_1 - Sepal.Width;
- X_2 - Petal.Width;
- X_3 - Petal.Length;
- X_4 - Sepal.Length.

The histogram units for each of these four variables are displayed in the Tables 4.1, 4.2, 4.3, 4.4.

Table 4.1: Histogram data for the variable X_1, Sepal.Width, of the Iris data set.

	X_1=Sepal.Width				
	$[2, 2.5[$	$[2.5, 3[$	$[3, 3.4[$	$[3.4, 3.9[$	$[3.9, 4.4]$
x_{11}	0.7	0.3	0.0	0.0	0.0
x_{21}	0.0	0.4	0.6	0.0	0.0
x_{31}	0.1	0.7	0.2	0.0	0.0
x_{41}	0.0	0.0	0.5	0.5	0.0
x_{51}	0.1	0.7	0.2	0.0	0.0
x_{61}	0.0	0.0	0.0	0.5	0.5
x_{71}	0.0	0.3	0.4	0.3	0.0
x_{81}	0.1	0.0	0.8	0.1	0.0
x_{91}	0.0	0.0	1.0	0.0	0.0
$x_{10\,1}$	1.0	0.4	0.5	0.0	0.0

Table 4.2: Histogram data for the variable X_2, Petal.Width, of the Iris data set.

	X_2=Petal.Width				
	$[0.1, 0.6[$	$[0.6, 1.1[$	$[1.1, 1.5[$	$[1.5, 2[$	$[2, 2.5]$
x_{12}	0.00	0.70	0.30	0.00	0.00
x_{22}	0.00	0.00	0.00	0.50	0.50
x_{32}	0.00	0.00	0.30	0.70	0.00
x_{42}	0.90	0.10	0.00	0.00	0.00
x_{52}	0.00	0.05	0.90	0.05	0.00
x_{62}	1.00	0.00	0.00	0.00	0.00
x_{72}	0.00	0.00	0.00	0.50	0.50
x_{82}	1.00	0.00	0.00	0.00	0.00
x_{92}	0.00	0.00	0.00	0.00	1.00
$x_{10\,2}$	0.00	0.00	0.90	0.10	0.00

Table 4.3: Histogram data for the variable X_3, Petal.Length, of the Iris data set.

	X_3=Petal.Length				
	$[1, 2.2[$	$[2.2, 3.3[$	$[3.3, 4.5[$	$[4.5, 5.7[$	$[5.7, 6.9]$
x_{13}	0.0	0.4	0.6	0.0	0.0
x_{23}	0.0	0.0	0.0	0.9	0.1
x_{33}	0.0	0.0	0.1	0.9	0.0
x_{43}	1.0	0.0	0.0	0.0	0.0
x_{53}	0.0	0.0	1.0	0.0	0.0
x_{63}	1.0	0.0	0.0	0.0	0.0
x_{73}	0.0	0.0	0.0	0.0	1.0
x_{83}	1.0	0.0	0.0	0.0	0.0
x_{93}	0.0	0.0	0.0	0.7	0.3
$x_{10\,3}$	0.0	0.0	0.4	0.6	0.0

Table 4.4: Histogram data for the variable X_4, Sepal.Length, of the Iris data set.

	X_4=Sepall.Length				
	$[4.3, 5[$	$[5, 5.7[$	$[5.7, 6.5[$	$[6.5, 7.2[$	$[7.2, 7.9]$
x_{14}	0.4	0.6	0.0	0.0	0.0
x_{24}	0.0	0.0	0.4	0.6	0.1
x_{34}	0.0	0.1	0.9	0.0	0.0
x_{44}	0.4	0.6	0.0	0.0	0.0
x_{54}	0.1	0.6	0.3	0.0	0.0
x_{64}	0.0	0.9	0.1	0.0	0.0
x_{74}	0.0	0.0	0.0	0.1	0.9
x_{84}	0.9	0.1	0.0	0.0	0.0
x_{94}	0.0	0.0	0.4	0.6	0.0
$x_{10\,4}$	0.0	0.0	0.6	0.4	0.0

Next, the several definitions of sample symbolic mean ($\bar{x}$), variance (s_1^2 and s_2^2) and covariance (cov_1 and cov_2), which were described previously for histogram variables, are applied to this data set. Each of these measures is estimated for two cases: in the first, none of the variables goes through the harmonization procedure; in the second, the harmonization procedure is performed on all the variables. By doing this, we can determine if the harmonization procedure has any effect on the computation of any of these measures. The values of the conventional sample mean, variance and covariance for the original single-valued data set are also presented for comparison purposes. All operations were implemented with the use of the R software [20] and are available upon request.

The definition of sample symbolic mean that was computed for the symbolic data set was the one presented in (4.3). The harmonization of the histograms was implemented, so that all the observations of X_1, $(x_{11}, .., x_{10\,1})$, were harmonized together with the ones of X_2, $(x_{12}, .., x_{10\,2})$, and all the observations of X_3, $(x_{13}, .., x_{10\,3})$, with the ones of X_4, $(x_{14}, .., x_{10\,4})$. For the single-valued data set, the definition of sample mean computed was the conventional one: for n observations x_i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The results obtained for both the symbolic (for the original and harmonized variables) and conventional

sample means are displayed in Table 4.5. By observing its results, it is possible to conclude that, for the computation of the sample symbolic mean of a histogram variable, the harmonization process has no effect on the results. Also, the values obtained for the sample mean are very similar for the histogram and the single-valued variables, which makes us assume that the sample symbolic mean definition for histogram variables is a good estimator of this measure, at least in this case.

Table 4.5: Sample means of the original histograms, harmonized histograms, and single-valued data.

	Original hist.	Harmonized hist.	Single values
$\bar{x}_1$	3.0600	3.0600	3.0573
$\bar{x}_2$	1.2595	1.2595	1.1993
$\bar{x}_3$	3.8720	3.872	3.7580
$\bar{x}_4$	5.8795	5.8795	5.8433

The results obtained for the computation of the sample variances are displayed in Table 4.6. The definitions of sample symbolic variances used were the ones presented in (4.6) and (4.9) (s_1^2 and s_2^2, respectively). The harmonization procedure followed the same logic as for the previous simulation of the sample symbolic mean. For the single-valued data set, the definition of sample variance used was the conventional one: for n observations, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where $\bar{x}$ is the conventional definition of sample mean. Through the observation of Table 4.6, it is possible to infer, once again, that there is no difference between the original and harmonized histogram variables for the values of s_1^2 and s_2^2. It can also be assumed that both s_1^2 and s_2^2 are good symbolic estimators for this case, since the values obtained for the histogram variables are relatively close to the ones obtained for the single-valued data.

Table 4.6: Sample variances of the original histograms, harmonized histograms and single-valued data.

	Original hist.	Harmonized hist.	Single values
$s_1^2(X_1)$	0.2369	0.2369	0.1900
$s_1^2(X_2)$	0.5143	0.5143	0.5810
$s_1^2(X_3)$	3.0111	3.0111	3.1163
$s_1^2(X_4)$	0.8220	0.8220	0.6857
$s_2^2(X_1)$	0.1491	0.1491	0.1900
$s_2^2(X_2)$	0.4681	0.4681	0.5810
$s_2^2(X_3)$	2.7694	2.7694	3.1163
$s_2^2(X_4)$	0.6755	0.6755	0.6857

The results attained for the sample symbolic covariances are presented in Table 4.7. The definitions of sample symbolic covariance used were the ones presented in (4.15) and (4.18) (cov_1 and cov_2, respectively). This time, the harmonization was done for all the observations of the two histograms variables for which the sample covariance was being calculated. For instance, to compute $cov_1(X_1, X_2)$, all the observations of X_1, $(x_{11}, .., x_{10\,1})$, were harmonized together with the ones of X_2, $(x_{12}, .., x_{10\,2})$. In the cases where the histograms for which the sample symbolic covariance was being computed were the same, the harmonization was made in a similar way to what was previously described for the sample symbolic mean and variances. By inspecting the results obtained, one can notice that, for cov_1, they are the same whether a histogram is harmonized or not. However, this does not happen for cov_2, where the two different procedures led to slightly different results. This means that, in the computation of cov_2, it matters if the histograms are harmonized or not. It can also be observed that, when the sample sym-

bolic covariance is being computed for two equal histograms, in the case of cov_1, it generates values equal to the ones obtained previously for s_2^2, as expected. For the estimation of the sample covariance in the single-valued data, the conventional definition of sample covariance was used, such that: $cov(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$. By comparing the results obtained using this conventional sample covariance for single-valued data with the ones from the sample symbolic covariances for histograms, it can be seen that there occurs a discrepancy of the signs only for the sample covariance between X_1 and X_4. However, the values are so close to zero in both cases, that it should not be very relevant. For all the other cases, both cov_1 and cov_2 seem to be good symbolic estimators of the covariance in this data set.

Table 4.7: Value of the covariances of the original/harmonized histograms and single-valued data.

	Original hist.	Harmonized hist.	Single values
$cov_1(X_1, X_1)$	0.1491	0.1491	0.1900
$cov_1(X_2, X_2)$	0.4681	0.4681	0.5810
$cov_1(X_3, X_3)$	2.7694	2.7694	3.1163
$cov_1(X_4, X_4)$	0.6755	0.6755	0.6857
$cov_1(X_1, X_2)$	-0.0702	-0.0702	-0.1216
$cov_1(X_1, X_3)$	-0.2102	-0.2102	-0.3296
$cov_1(X_1, X_4)$	0.0267	0.0267	-0.0424
$cov_1(X_2, X_3)$	1.1032	1.1032	1.2956
$cov_1(X_2, X_4)$	0.4813	0.4813	0.5163
$cov_1(X_3, X_4)$	1.2251	1.2251	1.2743
$cov_2(X_1, X_1)$	0.2053	0.1879	0.1900
$cov_2(X_2, X_2)$	0.4925	0.4767	0.5810
$cov_2(X_3, X_3)$	2.9251	2.8370	3.1163
$cov_2(X_4, X_4)$	0.7513	0.7182	0.6857
$cov_2(X_1, X_2)$	-0.0541	-0.0546	-0.1216
$cov_2(X_1, X_3)$	-0.1561	-0.1535	-0.3296
$cov_2(X_1, X_4)$	0.0635	0.0610	-0.0424
$cov_2(X_2, X_3)$	1.1673	1.1285	1.2956
$cov_2(X_2, X_4)$	0.5233	0.4981	0.5163
$cov_2(X_3, X_4)$	1.3331	1.2819	1.2743

From the previous example, it is possible to conclude that the harmonization of a histogram only affects the results of the second definition of sample symbolic covariance presented in this work, cov_2. For all the other measures, the harmonization procedure seems to be irrelevant to the results obtained. Through the comparison with the corresponding measures for the single-valued data set, all these sample symbolic measures for histograms appear to be good estimators, since they gave relatively similar results to their single-valued counterparts.

Chapter 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 presented in Chapter 3, to propose a new estimation method for SPCA, when it is applied to histogram-valued data. In the first place, a short overview of the conventional Principal Components Analysis (PCA) and also of some relevant cases of SPCA for histograms is provided. This is followed by the explanation of our proposal to improve the current SPCA methods that deal with histogram-valued variables, through the computation of linear combinations of histograms. Finally, this method is applied to two data sets, for three definitions of covariance(correlation) matrices.

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. This statistical method is used in exploratory data analysis and also as a method to reduce the dimension of the data.

The first PC generated through this method is defined as the linear combination of the original variables with highest variance, the second PC is the linear combination of the original variables non-correlated with the first PC with the highest variance, and so on. 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

$$\boldsymbol{PC}_l = \boldsymbol{X}\boldsymbol{\delta}_l. \tag{5.1}$$

This operation is easily done for single-valued data, but when it comes to symbolic data, the case is somewhat more complex and the best way to perform Symbolic PCA (SPCA) on symbolic variables is still under research and far from being fully developed.

Since the result of the different definitions of symbolic covariance is always given by a single-value, as presented in Chapter 4, one of the main challenges in SPCA is to define the linear combination of the original symbolical variables while using the single-valued elements of the eigenvectors $\boldsymbol{\delta}_l$ as weights, in such a way that the outcome is also a symbolic variable. This approach is called symbolic-conventional-symbolic. There is also an approach where the values of the covariances are presented in symbolic form, which is called symbolic-symbolic-symbolic, but that approach is not considered in this work.

In [13], an expression for the SPCA of interval-valued variables was generalized, generating PCs which are also interval variables. Considering that the original interval data is represented by a matrix $\boldsymbol{Y}$, which can also be represented by its respective matrices of the centers ($\boldsymbol{C}$) and ranges ($\boldsymbol{R}$), where the columns of these matrices are associated to the variables and the rows to the observations, the expression for the vector of the scores of the l-th interval Principal Component, $\boldsymbol{PC}_l$, is

$$\boldsymbol{PC}_l = [\boldsymbol{C}\boldsymbol{\delta}_l - \frac{1}{2}\boldsymbol{R}|\boldsymbol{\delta}_l|, \boldsymbol{C}\boldsymbol{\delta}_l + \frac{1}{2}\boldsymbol{R}|\boldsymbol{\delta}_l|]. \tag{5.2}$$

This expression is similar to the expression for Moore's Interval Algebra in (3.1).

In works where the SPCA is applied to histogram-valued variables, several different ways to compute the PC scores are used. However, in none of them are the PC scores defined as the linear combination of histograms, as it is proposed in this work. Before proceeding with the explanation of our proposal, 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. Nonetheless, in our work we do not consider the cases where the $\boldsymbol{\delta}_l$s originate from a covariance(correlation) matrix where the corresponding covariances are calculated only through the relationship between the probabilities of the subintervals (see [21]), and the contribution of the centers and/or ranges of the subintervals is ignored. This is due to the fact that, for our proposed estimation method, we consider that the information contained in the subintervals, and not only in their associated probabilities, is essential to obtain the best results.

A covariance matrix, represented as $\boldsymbol{\Sigma}$, is a collection of all the covariances between the k variables of a data set, such that $\boldsymbol{\Sigma}_{ij} = cov(X_i, X_j)$. By analyzing several other works about SPCA applied to histogram variables, one can conclude that several different covariance(correlation) matrices are used for the calculation of the eigenvectors $\boldsymbol{\delta}_l$. Some of them are not even directly related to the centers and ranges of the original subintervals, as it happened for the cov_1 and cov_2 defined in Chapter 4, but are connected to quantiles or even only to the probabilities of the subintervals. The authors, articles and corresponding definition of the sample symbolic covariance(correlation) matrix for histograms, which are used here, are the following:

- M. Kallyth, E.Diday in [19] and M. Chen, H. Wang in [22]:

  The definition of sample symbolic covariance used in both works is the same as cov_1, which was

previously defined according to the expression in (4.15). Thus, the sample symbolic covariance matrix obtained through this method, which is denoted by $\mathbf{\Sigma}^{(1)}$ in this work, is given by

$$\mathbf{\Sigma}^{(1)} = \begin{bmatrix} cov_1(X_1, X_1) & \dots & cov_1(X_1, X_j) & \dots & cov_1(X_1, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov_1(X_l, X_1) & \dots & cov_1(X_l, X_l) & \dots & cov_1(X_l, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov_1(X_p, X_1) & \dots & cov_1(X_p, X_j) & \dots & cov_1(X_k, X_p) \end{bmatrix}.$$

The eigenvectors obtained from this covariance matrix are then used to calculate the PCs, which are afterwards represented either as hypercubes or as intervals lengths in [19] and with a probability density function in [22].

- J. Le-Rademacher and L.Billard in [23]:
  Here the definition of sample symbolic covariance is equal to the previously defined cov_2 in (4.17). This definition of covariance is used to build the sample symbolic covariance matrix, which is denoted by $\mathbf{\Sigma}^{(2)}$:

$$\mathbf{\Sigma}^{(2)} = \begin{bmatrix} cov_2(X_1, X_1) & \dots & cov_2(X_1, X_j) & \dots & cov_2(X_1, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov_2(X_l, X_1) & \dots & cov_2(X_l, X_l) & \dots & cov_2(X_l, X_p) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ cov_2(X_p, X_1) & \dots & cov_2(X_p, X_j) & \dots & cov_2(X_k, X_p) \end{bmatrix}.$$

  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 PC scores as histograms. However, this polytope method can be somewhat complex and hard to interpret.

- M. Ichino in [24]:
  This SPCA method for histogram variables starts by computing quantiles from the quantile function of the histograms, as previously presented in (2.13). 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

$$Q_{ijl} = Q_{x_{jl}}(\frac{i}{d}), \tag{5.3}$$

  with $i = \{0, 1, ..., d\}$, $j = \{1, ..., k\}$, $l = \{1, ..., p\}$.

  After having created these quantiles for all the variables and observations in the data, 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 $\boldsymbol{Q}$, so that:

$$\boldsymbol{Q} = \begin{bmatrix} Q_{011} & \cdots & Q_{01l} & \cdots & Q_{01p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{d11} & \cdots & Q_{d1l} & \cdots & Q_{d1p} \\ Q_{021} & \cdots & Q_{02l} & \cdots & Q_{02p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{d21} & \cdots & Q_{d2l} & \cdots & Q_{d2p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{0k1} & \cdots & Q_{0kl} & \cdots & Q_{0kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Q_{dk1} & \cdots & Q_{dkl} & \cdots & Q_{dkp} \end{bmatrix}.$$

Afterwards, a sample symbolic correlation matrix is computed for the data table presented in matrix $\boldsymbol{Q}$ using either the Spearman's or the Pearson's correlation coefficient. The Spearman's correlation coefficient, ρ, is given by the expression

$$\rho = 1 - \frac{6 \sum d_i^2}{k(k^2 - 1)}, \tag{5.4}$$

where k is the number of observations and d_i is the difference between the ranks of corresponding variables. The Pearson's correlation coefficient, r_{qz}, is given by

$$r_{qz} = \frac{(k \times (d+1)) \sum q_i z_i - \sum q_i \sum z_i}{\sqrt{(k \times (d+1)) \sum q_i^2 - (\sum q_i)^2} \sqrt{(k \times (d+1)) \sum z_i^2 - (\sum z_i)^2}}, \tag{5.5}$$

where q_i and z_i are the i-th elements from the columns of $\boldsymbol{Q}$ that are related to the variables between which we want to compute the corresponding correlation.

The results obtained through this method are then represented as a series of d arrow lines for each observation. The sample symbolic correlation matrix used in our work, based on this method, is denoted by $\boldsymbol{\Sigma}^{(3)}$ and only the Spearman's correlation coefficient is used for its computation, as it gave better results in the example from [24].

In [25], a method similar to this one is presented, but as it uses more complex correlation coefficients, it was not considered to be worth studying for our purposes.

5.2 SPCA using linear combinations of histograms

As it was previously seen, in most of the works about SPCA applied to histogram-valued data, the PCs obtained are not represented as histogram-valued variables (see [21], [19], [22], [24], and [25]). As PCA is very commonly used as a technique to reduce the dimension of the original data, by selecting the first few PCs to be used in other statistical methods, it is more useful if the resulting PCs are also histogram variables.

To improve the current SPCA methods that deal with histogram-valued variables, it is shown in this

work how to obtain histogram-valued scores, by following the same idea of (5.2). However, instead of using an expression related to Moore's Interval Algebra, we apply the new algebra for histograms, previously introduced in Chapter 3, to compute the linear combinations of histograms.

This SPCA method is applied to a data set X with p histogram-valued variables $X_l = (\boldsymbol{C}_l, \boldsymbol{R}_l, \boldsymbol{P}_l)$, with $l = \{1, .., p\}$, and k observations. Each observation j, $j = \{1, .., k\}$, corresponds to a histogram with a number of subintervals for the variable l equal to n_{jl}, which may vary. Each of these subintervals is associated to the index i, $i = \{1, .., n_{jl}\}$. Thus, for instance, c_{ijl} corresponds to the center of the subinterval i of the observation j associated to the histogram variable X_l. Before proceeding, it is necessary to define some new matrices for the centers, ranges, and associated probabilities of the histograms. Unlike the $\boldsymbol{C}_l$, $\boldsymbol{R}_l$, $\boldsymbol{P}_l$ matrices used in Chapter 4, where each matrix aggregated the information from all the observations of a variable l, these new matrices combine all the information of an observation j, for all the p variables of the 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. Thus, the matrices $\boldsymbol{C}_j$, $\boldsymbol{R}_j$ and $\boldsymbol{P}_j$ are defined, such that:

$$\boldsymbol{C}_j = \begin{bmatrix} c_{1j1} & \cdots & c_{1jl} & \cdots & c_{1jp} \\ c_{2j1} & \cdots & c_{2jl} & \cdots & c_{2jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{ij1} & \cdots & c_{ijl} & \cdots & c_{ijp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{n_{j1}j1} & \cdots & c_{n_{j1}jl} & \cdots & c_{n_{j1}jp} \\ \vdots & \ddots & \vdots & \ldots & \vdots \end{bmatrix}, \boldsymbol{R}_j = \begin{bmatrix} r_{1j1} & \cdots & r_{1jl} & \cdots & r_{1jp} \\ r_{2j1} & \cdots & r_{2jl} & \cdots & r_{2jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{ij1} & \cdots & r_{ijl} & \cdots & r_{ijp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{n_{j1}j1} & \cdots & r_{n_{j1}jl} & \cdots & r_{n_{j1}jp} \\ \vdots & \ddots & \vdots & \ldots & \vdots \end{bmatrix},$$

$$\boldsymbol{P}_j = \begin{bmatrix} p_{1j1} & \cdots & p_{1jl} & \cdots & p_{1jp} \\ p_{2j1} & \cdots & p_{2jl} & \cdots & p_{2jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{ij1} & \cdots & p_{ijl} & \cdots & p_{ijp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n_{j1}j1} & \cdots & p_{n_{j1}jl} & \cdots & p_{n_{j1}jp} \\ \vdots & \ddots & \vdots & \ldots & \vdots \end{bmatrix}.$$

A probability of zero is associated with the elements c_{ijl} and r_{ijl} of these matrices where $i > n_{jl}$. Consequently, their value is not relevant and can be chosen arbitrarily. 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 (3.25) and (3.26), 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 to this situation the general expression for the linear combinations from the algebra based on the operations with quantile functions presented in (3.29), each observation j of the l-th PC (score), PC_{jl} is

$$PC_{jl} = (\boldsymbol{C}_j^{\bullet}|\boldsymbol{\delta}_l| \pm \frac{1}{2}\boldsymbol{R}_j^{\bullet}|\boldsymbol{\delta}_l|, \boldsymbol{P}_{j1}^{\bullet}), \tag{5.6}$$

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 Principal Components $\boldsymbol{PC}_l$.

Now that a way to compute histogram PCs was defined, the only thing left to do is to determine the definition of covariance(correlation) that should be used for the construction of the covariance(correlation) matrix from which the eigenvectors $\boldsymbol{\delta}_l$ are computed. In the following examples, the previously defined $\boldsymbol{\Sigma}^{(1)}$, $\boldsymbol{\Sigma}^{(2)}$, and $\boldsymbol{\Sigma}^{(3)}$ are used in order to verify the impact of each definition on the results obtained.

5.3 Examples

The final step of this work consisted in implementing the expression (5.3) 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 sample symbolic covariance(correlation) matrices, $\boldsymbol{\Sigma}^{(1)}$, $\boldsymbol{\Sigma}^{(2)}$ and $\boldsymbol{\Sigma}^{(3)}$, defined previously. Two examples, based on real data sets were explored using SPCA: Iris flower [19] and Hardwood [26]. Based on the proposed definition of linear combinations of histograms, the associated scores were obtained. All the simulations presented next were done with the use of the R software [20]. The functions implemented on R are available upon request.

5.3.1 Iris data set

The first data set where this method was implemented is the histogram-valued Iris flower data set presented previously in Chapter 4, whose four variables are represented in the Tables 4.1, 4.2, 4.3, and 4.4. The four histogram variables of this data describe the following features of each Iris flower:

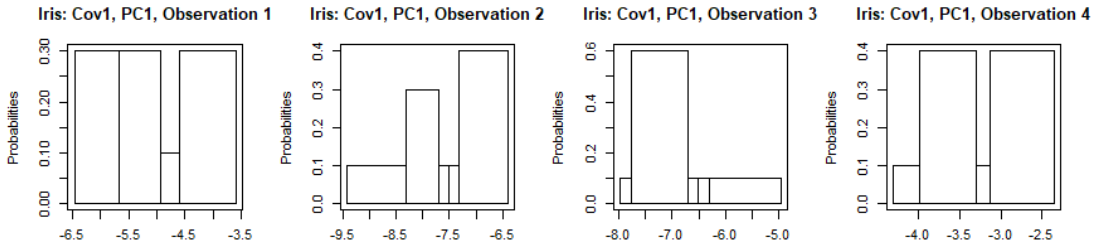
- X_1 - sepal width;
- X_2 - petal width;
- X_3 - petal length;
- X_4 - sepal length.

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

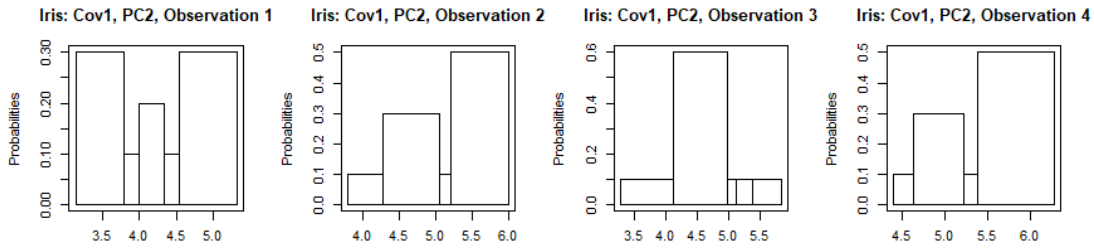
$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 0.149 & -0.070 & -0.210 & 0.027 \\ -0.070 & 0.468 & 1.103 & 0.481 \\ -0.210 & 1.103 & 2.769 & 1.225 \\ 0.027 & 0.481 & 1.225 & 0.675 \end{bmatrix}, \boldsymbol{\delta}^{(1)} = \begin{bmatrix} 0.053 & 0.742 & -0.395 & 0.539 \\ -0.342 & -0.069 & -0.813 & -0.467 \\ -0.854 & -0.214 & 0.112 & 0.461 \\ -0.389 & 0.631 & 0.413 & -0.528 \end{bmatrix}.$$

The eigenvectors correspond to the four columns of $\boldsymbol{\delta}^{(1)}$, which are used to obtain the symbolic PC scores through the expression (5.3). 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. Since both of 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. The histograms for the first four observations of the first and second PCs are represented in Figure 5.1. By analysing it, it is possible to observe that the distribution of the histograms for these four observations varies considerably for both the first and second PC. For the first PC, the observation that should have a lower value of Petal length is observation 4, because the corresponding histogram has the subintervals with the higher values for the bounds. For the remaining observations, it is harder to make comparisons. For the second PC, one can conclude, for example, that the observations 2 and 4 have very similar values of sepal length and width, as their respective histograms are very much alike. However, it is hard to find other patterns in the data when the results are represented in this way.



(a) PC_1 scores



(b) PC_2 scores

Figure 5.1: First and second PCs (scores of obs. 1 to 4) when cov_1 is applied to the Iris data set.

To overcome the aforementioned 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 5.2. The different number of rectangles among the several observations is due to the fact that each observation goes through a harmonization procedure for the computation of both the $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$, which can create a different number of subintervals for each case. Through the analysis of Figure 5.2, it can be concluded that there is considerable variability of the distribution of the histograms among all the observations. This is consistent with the fact that these observations were obtained through a K-means method, which creates partitions with the highest variability possible between clusters (in this case between the micro-data summarized in the form of a histogram). However, now it is easier to find patterns in the data and reach some conclusions. Therefore, observation 5 seems to have been created by the aggregation of flowers which had 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 which had 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.

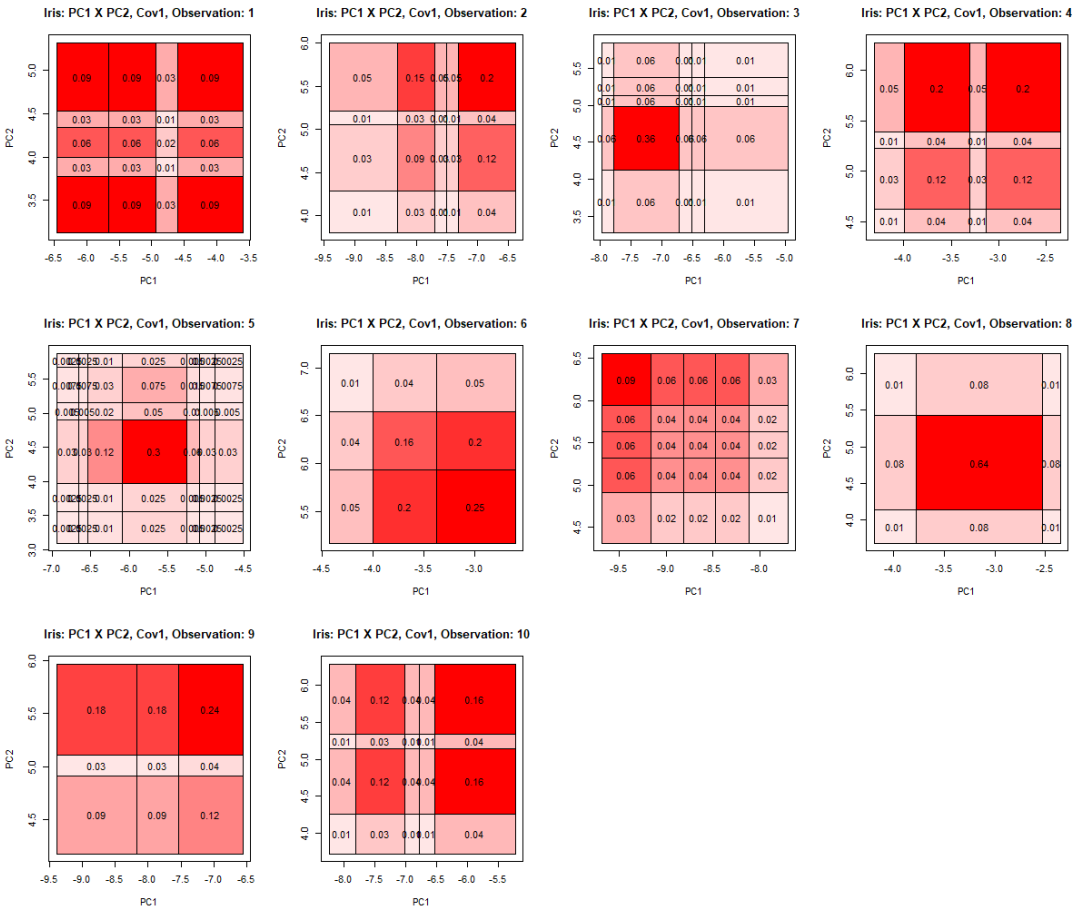


Figure 5.2: Joint probability of PC_1 and PC_2 scores when cov_1 is applied to the Iris data set.

It is also interesting to check what the sample symbolic covariance matrix for the PCs is, when the

definition of sample symbolic covariance cov_1 is used (computation based on the obtained scores). This matrix, represented as $\mathbf{\Sigma}_{PC}^{(1)}$, is the following:

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

The values of the main diagonal of $\mathbf{\Sigma}_{PC}^{(1)}$ 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.

If, instead of using cov_1 to obtain the covariance matrix in the iris data set, the definition of sample covariance cov_2 is used, the sample covariance matrix $\mathbf{\Sigma}^{(2)}$ and corresponding eigenvectors $\boldsymbol{\delta}^{(2)}$ are

$$\mathbf{\Sigma}^{(2)} = \begin{bmatrix} 0.205 & -0.054 & -0.156 & 0.063 \\ -0.050 & 0.492 & 1.167 & 0.523 \\ -0.156 & 1.167 & 2.925 & 1.333 \\ 0.063 & 0.523 & 1.333 & 0.751 \end{bmatrix}, \boldsymbol{\delta}^{(2)} = \begin{bmatrix} 0.033 & 0.804 & 0.485 & -0.341 \\ -0.341 & -0.080 & 0.639 & 0.685 \\ -0.850 & -0.197 & 0.043 & -0.486 \\ -0.400 & 0.554 & -0.596 & 0.421 \end{bmatrix}.$$

By applying the columns of $\boldsymbol{\delta}^{(2)}$ in the expression (5.3) and, thus, obtaining the PCs for this case, 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 $\boldsymbol{\delta}^{(2)}$ are very similar, which means that the first and second PCs obtained for both cases have a similar interpretation and lead to similar scores. Therefore, the same reasoning that was applied for the previous case can also be applied here. Accordingly, the joint probability graph of $\boldsymbol{PC}_1$ and $\boldsymbol{PC}_2$ for cov_2 (see Figure 5.3) is very similar to the one obtained for cov_1 in Figure 5.2, and the same conclusions can be reached.

Next, by calculating the sample covariance matrix for the PCs reached through cov_2, $\mathbf{\Sigma}_{PC}^{(2)}$, 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, when computing $cov_2(X_1, X_2)$ and $X_1 = X_2$, none of the variances presented in Chapter 4 is obtained. Hence, this value should not be 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, if the sample correlation matrix introduced by M. Ichino in [24], $\mathbf{\Sigma}^{(3)}$, and the corresponding

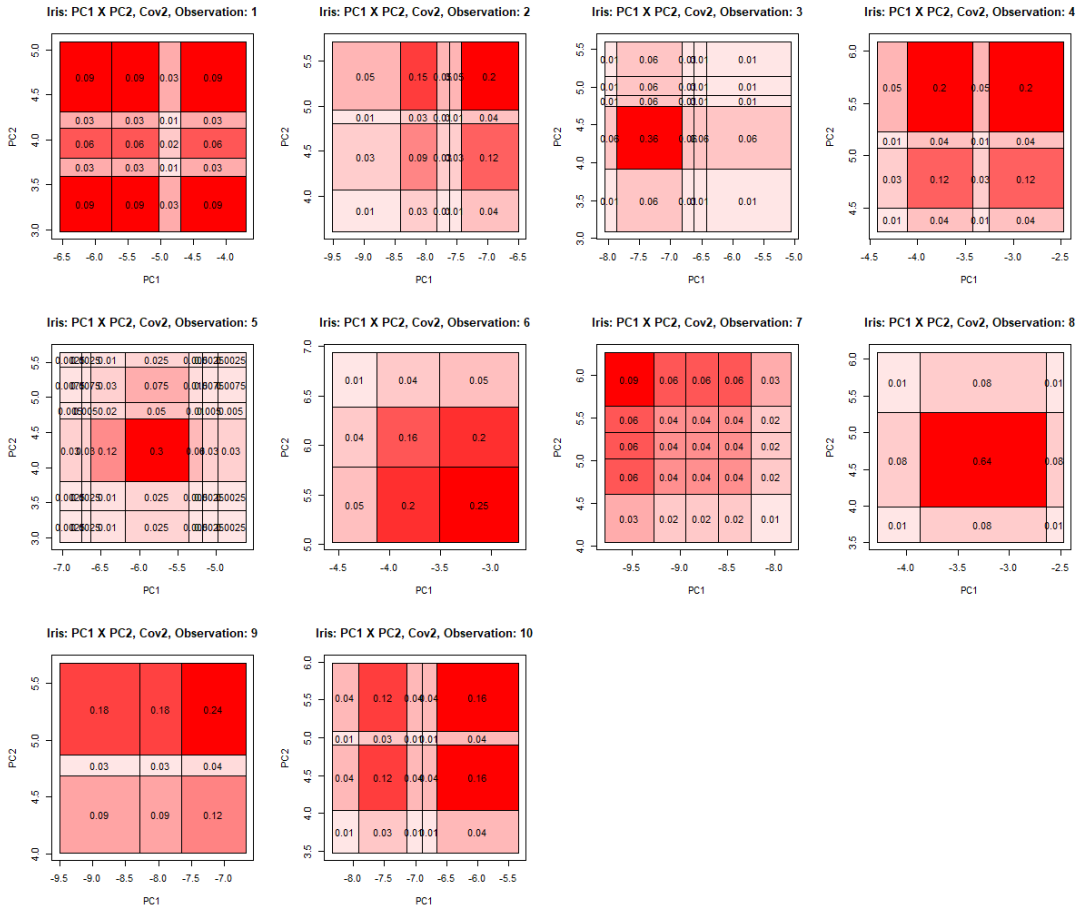


Figure 5.3: Joint probability of PC_1 and PC_2 scores when cov_2 is applied to the Iris data set.

eigenvectors, $\boldsymbol{\delta}^{(3)}$, are computed for the Iris data set, by considering $d = 4$ quantiles and the Spearman's correlation coefficient, the outcome is the following:

$$\boldsymbol{\Sigma}^{(3)} = \begin{bmatrix} 1 & 0.131 & 0.073 & 0.402 \\ 0.131 & 1 & 0.955 & 0.865 \\ 0.073 & 0.955 & 1 & 0.892 \\ 0.402 & 0.865 & 0.892 & 1 \end{bmatrix}, \boldsymbol{\delta}^{(3)} = \begin{bmatrix} -0.184 & 0.944 & -0.219 & 0.164 \\ -0.565 & -0.188 & -0.669 & -0.444 \\ -0.567 & -0.242 & 0.025 & 0.787 \\ -0.570 & 0.122 & 0.709 & -0.395 \end{bmatrix}.$$

All the columns of $\boldsymbol{\delta}^{(3)}$ are significantly different when comparing them with the two previous cases. By applying these eigenvectors to build the symbolic PCs, 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. The second PC explains 25.18% of the variance, the third PC 2.27% and the fourth 0.7%. This time, the variables X_2, X_3, and X_4 are influential in a very similar proportion in the definition of $\boldsymbol{PC}_1$, while, for the definition of $\boldsymbol{PC}_2$, X_1 is by far the most relevant variable. The graph for the joint probabilities of the first and second PCs is represented in Figure 5.4. Despite the differences in the eignenvectors when comparing them with the two previous cases, the distribution of the histograms continues to be very similar. The only difference occurs in the bounds of the subintervals of the histograms, which seem to have been translated to the left in both the first and second PCs, when

comparing them with the graphs in Figures 5.2 and 5.3.

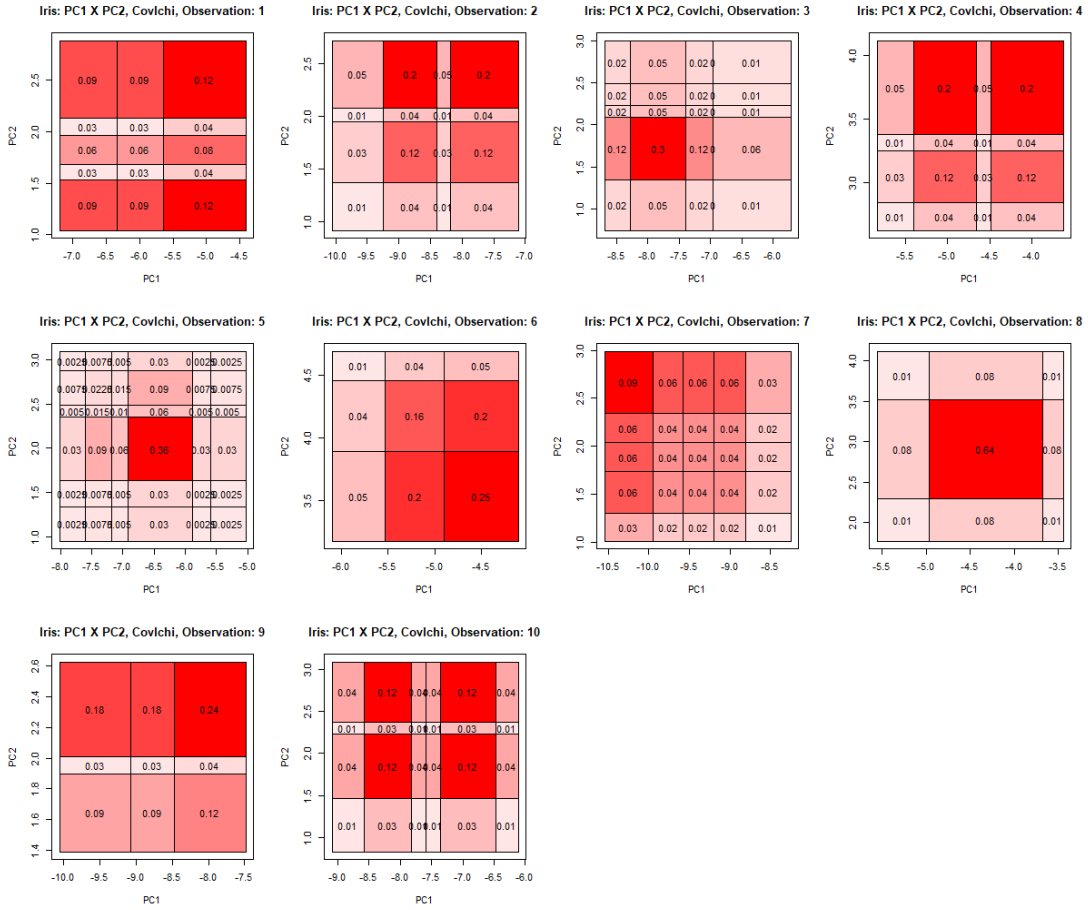


Figure 5.4: Joint probability between PC_1 and PC_2 scores when Ichino's correlation matrix is applied to the Iris data set.

When determining the sample correlation matrix through the same method for all these PCs, $\mathbf{\Sigma}_{PC}^{(3)}$, we get

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

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 $\Sigma^{(3)}$.

By summing up the results obtained for each of the three covariance(correlation) matrices, it is possible to infer that, through the analysis of the joint probability graphs for the first and second PCs, one can find some patterns in the data that can be useful for its study. However, a visual examination of these graphs is not enough to make a deeper analysis of the data, and other statistical methods need to be applied to fully analyze it. The distribution of the PCs scores, for $\boldsymbol{PC_1}$ and $\boldsymbol{PC_2}$, seems to be very similar

in the three methods. Nonetheless, the results seem to be more trustworthy for the PCs obtained with cov_1, as it was possible to verify, through that same measure, that the PCs obtained are uncorrelated. For the two other methods, the values of the covariance(correlation) measure did not equal zero when the PCs were different. In SPCA, the principal components are expected to be uncorrelated with one another. Thus, for our proposed estimation method, cov_1 is the best option of the three tested.

5.3.2 Hardwood data set

The second data set to which the method developed in this work to perform histogram SPCA was applied, was obtained from a US Geological Survey of Hardwood Trees that can be found in [26]. 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 $5^{\text{o}}C \times 1000$;
- X_8: Moisture index.

All the observations of these variables were already represented as histograms in the original data set. However, before applying the SPCA, it is necessary to standardize all of these variables because they are originally presented in different scales, which could lead to misleading results. Each standardized histogram observation j of the eight variables X_l, $\widetilde{x_{jl}}$, is obtained by applying the formula $\widetilde{x_{jl}} = \frac{x_{jl}-\bar{x}_l}{\sqrt{s^2(X_l)}}$, where $\bar{x}_l$ is the sample symbolic mean in (4.3), and the definition of sample symbolic variance used to compute $s^2(X_l)$ varies with each method: s_2^2 in (4.9) when we compute $\mathbf{\Sigma}^{(1)}$ and s_1^2 in (4.6) when we compute $\mathbf{\Sigma}^{(2)}$ and $\mathbf{\Sigma}^{(3)}$.

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

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

In this case, a first PC, which explains 52.5% of the variance, and a second PC, which explains 27.1%, were obtained. The first PC 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 the growing degree days (a measure of heat accumulation), as they are the most relevant (in a negative proportion). For the second PC, the two most significant variables for its definition (also in a negative proportion) 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 5.5. It can be observed that there seem to be increasingly higher probabilities the closer you are to the center, even if the distribution is slightly different for each observation. This may be a consequence of the standardization of the data, which was performed beforehand. In cases like this, it can be hard to reach any relevant conclusion through the analysis of these graphs. The interpretation might have been easier if new graphs had been built, where the same scales had been used for the x and y axes for all the observations, and areas with a white color in the new graphs had been defined for the regions where the observations had an associated probability of 0. However, this was not tested for this work.

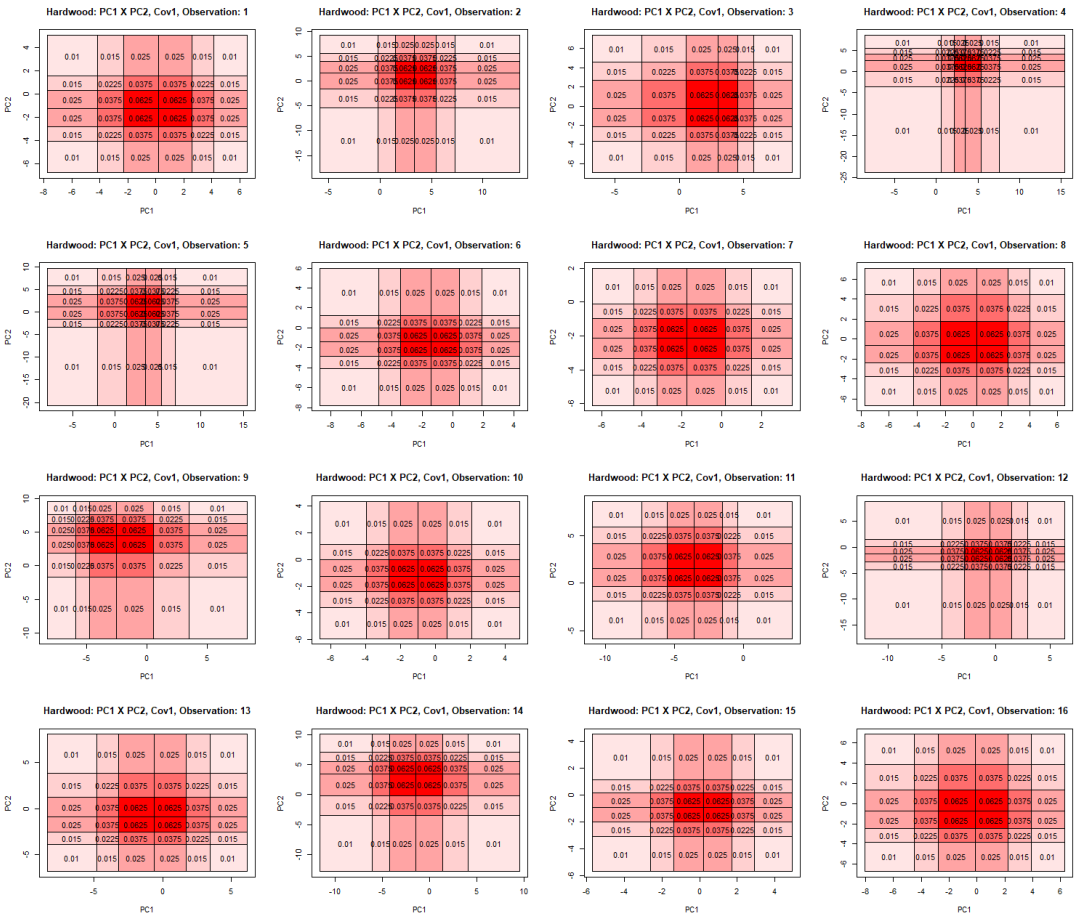


Figure 5.5: Joint probability between PC_1 and PC_2 scores when cov_1 is applied to the Hardwood data set.

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. Therefore, as the results for the PCs obtained through the other definitions of covariance(correlation) would add no interesting additional information, they were omitted from this work.

This case allows us to infer that sometimes it might be difficult to perform a visual analysis of the data only through a visualization of the joint probability graph of the first and second PCs. When that happens, other statistical methods and visualization tools need to be used to accomplish deeper analyzes.

Chapter 6

Conclusions and future work

6.1 Conclusions

This work had two main objectives: firstly, to describe an algebra for histograms based on the arithmetic operations with quantile functions introduced in [1]; secondly to apply it to the SPCA statistical, thus proposing a new estimation method.

Chapter 3 accounts for the achievement of finding the general expression (3.29) for the computation of linear combinations according to this histogram algebra, which allows us to easily perform operations with histogram-valued variables. This algebra for histograms follows a reasoning similar to that of Moore's Interval Algebra, which was presented in Chapter 2. In both of these algebras, 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. However, this is still the best option to perform arithmetic operations with intervals and histograms. If another algebra is used, where the ranges of the intervals/subintervals can also shrink in a linear combination, as it is the case of the Extended Interval Algebra, it creates degenerate intervals, which are even more problematic than the previous case.

There is also another aspect of this algebra with histograms that can create setbacks. The harmonization procedure explained in Chapter 3, which enables us to do operations with the quantile functions of the histograms, generates smaller and smaller subintervals when the cumulative probabilities associated to the histograms that take part in the operations are different. In most cases, this is not relevant, but in extreme cases where the probabilities used in the description of the histograms are very different, this can lead to very small subintervals, which can also lose their significance. Consequently, it is necessary to be careful and try to prevent these situations. A possible way to overcome this problem is by using subintervals with probabilities associated which are a multiple of 0.1, thus avoiding too many divisions of the subintervals when the harmonization is executed.

It was also checked if this algebra respected the eight axioms of a vector space and the conclusion was that it failed two of them: the existence of an additive inverse, and also the distributivity of scalar multiplication with respect to field addition. However, the remaining axioms hold and are useful, as it is

the associativity and commutativity of addition.

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 this area. The ones which also have histogram-valued PCs as the end result only achieve it through very complex methods, which are hard to interpret and can also be a waste of computational memory for very large data sets. In this work, we accomplished this through the linear combination of histogram variables using the expression (5.3), which enables a quick and easy computation of histogram-valued PCs.

This method was tested for three different definitions of covariance, two of which were described in Chapter 4 and were found to be good estimators of that measure for two data sets. 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 presented in (4.15), was it possible to verify that the Principal Components obtained were in fact uncorrelated, *i.e.*, that they had a symbolic covariance of 0, 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 visualization of the results was done through a joint probability graph of the histograms of the first and second PCs. This graph allows us to find some patterns in the data and to easily visualize their distribution for all the observations. However, for the second data set, it was hard to reach any relevant conclusions through the visualization of this graph and clearly some improvements are needed. Therefore, even if the observation of the joint probability graph 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 independent symbolic variables, which can also be helpful for many methods.

6.2 Future work

Several aspects of this work require further study and improvement, namely the following: the exploration of other properties of the histogram algebra hereby defined; the applicability of this algebra to other areas in which linear combinations with histograms are useful, besides SPCA, such as clustering, classification, regression analysis, among others.

Moreover, the joint probability graph obtained for the SPCA with the histogram algebra is sometimes hard to interpret, as it was seen in the Hardwood data set. Therefore, if a better way to visualize the results can be found, it can help us to more easily reach important conclusions regarding the data. It would also be relevant to check, for some symbolic statistical methods, if the results can be improved by using the corespondent data set with histogram-valued PCs, which could have its dimensions reduced by considering only the first few PCs that explain the vast majority of the variance.

Bibliography

[1] 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.

[2] P. Brito. Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4:281–295, 2014.

[3] K. Ishikawa. *Guide to Quality Control.* White Plains, New York: Kraus International Publications, 1982.

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

[5] C. Bertoluzza, N. Blanco, and A. Salas. *On a new class of distances between fuzzy numbers*. Mathware & Soft Computing, 1995.

[6] S. Dias. *Linear regression with empirical distributions*. PhD thesis, Faculdade de Ciências da Universidade do Porto, 2014.

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

[8] 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.

[9] A. Han, Y. Hong, and S. Wang. Autoregressive Conditional Models for Interval-Valued Time Series Data, December 2013.

[10] L. Stefanini. A generalization of Hukuhara difference and division for interval and fuzzy arithmetic. *Fuzzy Sets and Systems*, 161(11):1564 – 1584, 2010. ISSN 0165-0114. doi: https://doi.org/10.1016/j.fss.2009.06.009.

[11] A. Colombo and R. Jaarsma. A powerful numerical method to combine random variables. In *IEEE Transactions on Reliability*, pages 126–129. IEEE, 1980.

[12] L. Billard and E. Diday. From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98(462):470–487, 2003.

[13] M. R. Oliveira, M. Vilela, A. Pacheco, R. Valadas, and P. Salvador. Extracting Information from Interval Data Using Symbolic Principal Component Analysis. *Austrian Journal of Statistics*, 46 (3-4):79–87, Apr. 2017. doi: 10.17713/ajs.v46i3-4.673.

[14] L. Billard and E. Diday. Symbolic regression analysis. *Classification, Clustering and Data Analysis. Proceedings of the 8th Conference of the International Federation of Classification Societies (IFCS'02) Cracow, Poland, July 2002*, pages 281–288, 2002.

[15] F. de Carvalho, P. Brito, and H.-H. Bock. Dynamic clustering for interval data based on L_2 distance. *Computational Statistics*, 21(2):231–250, 2006.

[16] M. Vilela. Classical and Robust Symbolic Principal Component Analysis for Interval Data. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2015.

[17] L. Billard and E. Diday. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley Sons, Ltd., 2006.

[18] L. Billard. Sample covariance functions for complex quantitative data. *Proceedings of the 2008 World Conference International Association of Statistical Computing, Yokohama, Japan*, pages 157–163, 2008.

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

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

[21] O. Rodriguez, E. Diday, and S. Winsberg. Generalization of the principal components analysis to histogram data. *Proceedings 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases; Workshop on Symbolic Data Analysis*, 2000.

[22] 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.

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

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

[25] S. Makosso-Kallyth. Principal Axes Analysis of Symbolic Histogram Variable. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2:188–200, 2016.

[26] 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.