Multi-class Classification of Distributional Data

Ana Rodrigues dos Santos
Master in Data Science
Department of Computer Science
2023

Supervisor
Sónia Dias, Associate Professor at Instituto Politécnico de Viana do Castelo

Co-supervisor
Paula Brito, Associate Professor at Faculty of Economics University of Porto

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Sworn Statement

I, Ana Carolina Silva Rodrigues dos Santos, enrolled in the Master Degree in Data Science at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offence.

Ana Rodrigues dos Santos
Porto, August 25, 2023

" ...it's the magic of fighting battles beyond endurance, beyond cracked ribs, ruptured kidneys and detached retinas. It's the magic of risking everything for a dream that nobody sees but you. "

Morgan Freeman as *Eddie Dupris*, written by Paul Haggis

Acknowledgements

I would like to thank professors Sónia Dias and Paula Brito for the guidance and effort they put in. Their supervision was crucial for the work developed in this thesis. They were both valuable whether it was answering my questions or putting me in contact with people who would give me the answers that I needed. A special thank you to Paula Alexandra Amaral, professor at FCT UNL, is in order. She provided a precious help regarding the optimisation process.

To my family and friends, a huge thank you for the support and for just being themselves. Particularly, I wish to thank my father who always encourages me to go above and beyond without knowing or understanding what I actually do.

Resumo

Esta tese aborda o problema de classificação de dados distribucionais, onde cada unidade é descrita por variáveis histograma ou intervalares. O método proposto utiliza uma função linear discriminante que, sob hipóteses específicas, permite a representação de distribuições e intervalos com recurso a funções quantil. A função discriminante garante a definição de um *score* para cada unidade que, por sua vez, concede a possibilidade de classificar as referidas unidades em grupos *a priori*, utilizando a distância de Mallows. Esta dissertação tem como objetivo estender o método desenvolvido anteriormente, que permite a classificação em duas classes, para proporcionar métodos que permitam a classificação considerando mais do que duas classes *a priori*, fazendo uso de três mecanismos diferentes. Estas técnicas abrem caminho para o desenvolvimento de análises de dados simbólicos, uma vez que melhoram as ferramentas atuais utilizadas para a classificação binária.

A ilustração dos métodos propostos para a classificação multi classe é feita com recurso a três casos. Os dados reais utilizados estão associados ao tráfego de Internet e a modelos de carros. Os dados produzidos artificialmente estão associados a redes, no sentido de grafos. Mais ainda, a implementação foi desenvolvida em linguagem R.

Palavras Chave: Classificação Multi-classe, Análise de Dados Simbólicos, Dados Distribucionais, Função Discriminante Linear, Distância de Mallows

Abstract

This thesis addresses a classification problem of distributional data, where data units are described by histogram or interval-valued variables. The proposed method uses a linear discriminant function where, under specific assumptions, distributions or intervals are represented by quantile functions. The discriminant function grants the definition of a score for each unit which enables the classification of the units in *a priori* groups, using the Mallows distance. The aim of this dissertation is to extend the method previously developed for two-class classification, in order to provide the means to perform the classification with more than two *a priori* classes, using three distinct approaches. These approaches contribute to the development of Symbolic Data Analysis, by improving the current method for binary classification.

The illustration of the methods proposed for the multi-class classification is performed by using three cases. The real data sets are associated with Internet traffic and car models. The synthetic data sets are associated with networks, in the sense of graphs. All methods are implemented in R.

Keywords: Multi-class Classification, Symbolic Data Analysis, Distributional Data, Linear Discriminant Function, Mallows Distance

Contents

List of Tables

List of Figures

List of Listings

Acronyms

BGP	Border Gateway Protocol
BI	Between Inertia
CLDF	Consecutive Linear Discriminant Functions
LDA	Linear Discriminant Analysis
OVA	One-Versus-All
OVO	One-Versus-One
RTT	Round-Trip Time
SDA	Symbolic Data Analysis
TI	Total Inertia

Chapter 1

Introduction

Similarly to classical data, symbolic data may display an underlying structure that aggregates units together in classes. Hence, classifying distributional data in two or more *a priori* groups has become a prominent problem. This classification problem can be addressed in various ways. Linear Discriminant Analysis works by designing a linear combination of explanatory variables known as a discriminant function. This enables the definition of a score for each unit.

In this thesis, we propose a generalisation of the linear discriminant method for the binary classification of distributional data. Using a linear discriminant function, it is possible to develop a score function that is used to classify each unit in one of the *a priori* groups. The method for two classes has already been developed in [1], therefore the extension of this method to more than two classes is the main goal of this dissertation.

1.1 Motivation

We have been observing an increase in global data consumption. In fact, in 2020, when compared to the previous year, it has increased more than 30% [2]. This compels an improvement in storing data. The amount of data gathering pressures researchers and analysts to develop new strategies to efficiently examine data. One of these approaches lies in generalising the classical concepts of data sets. Symbolic Data Analysis's ambition is to be a viable tool to improve upon classical data analysis.

The method explored in this thesis assumes that the study is not based on an individual level, but at a group level. Hence, usually, the main idea is to aggregate classical data and apply a summary indicator such as median, mean or mode. Although this approach is simple and sounds reasonable, it incurs in a problem which is the loss of data variability. Variability translates how far apart data is, allowing us to determine how well we can generalise results. Therefore, losing this information may be critical. Apart from data aggregation, symbolic data can naturally be obtained from recording processes, representing underlying variability.

Overall, Symbolic Data Analysis (SDA) provides a framework to properly represent, understand and analyse data with variability that is explicitly considered. It is an emerging area of statistics where data may take the form, for example, of intervals or histograms. Examples of this data include the analysis of schools, when the data gathered concerns individual students, and analysis of cars' models (not specific vehicles). Nevertheless, with this approach, there are problems that arise: most existing concepts, methods and models are not appropriate for this data, since they are designed for single-valued observations. In order to fix this problem several methods have been adapted such as regression models ([3] and [4]) and likelihood-based inference [5].

1.2 Objective

The main goal of this thesis is to extend the currently existing linear discriminant models for interval or histogram-valued variables that allow for the prediction of histograms and, as a consequence, multi-class classification. We introduce and explore three techniques: One-Versus-One, One-Versus-All and Consecutive Linear Discriminant Functions. The first two methods assume that we are in possession of the baseline setting, i.e., the binary classification. The last method requires the definition and study of a symbolic correlation measure. The study of this measure also provides a new contribution to the development of Symbolic Data Analysis. Moreover, regarding Consecutive Linear Discriminant Functions, we studied two possible classification definitions.

1.3 Organisation of the thesis

The thesis is organised as follows.

- In Chapter 2 we introduce the main concepts useful to this work such as the type of symbolic variables, focusing especially on those that are considered in this study, histogram-valued variables. The concepts of quantile function, Mallows distance, barycentric histograms and inertia measure described in this chapter were developed for the original method. These are mainly discussed in [1], [6], [7] and [8]. It is also addressed the graphical representation of two histogram-valued variables. It is advanced definitions such as symbolic covariance, variance, standard deviation and linear correlation. Regarding the symbolic linear correlation, it is proven that some of the mathematical properties associated with the well known correlation coefficients for the classic data hold. Moreover, we explained how a rewriting operation may be performed, if needed. Finally, we describe the assumptions required to develop the method and the linear combination of symbolic variables.

- In Chapter 3 is introduced the linear discriminant function proposed in [1], allowing for the definition of a score associated with each unit. Moreover, the optimisation and classification processes for the case of two *a priori* classes is discussed. Finally, we present the three approaches that provide multi-class classification, describing both the optimisation and the classification procedures performed.

- In Chapter 4 the computational implementation details and the choices made, regarding both the processes for the case of two *a priori* classes and more than two *a priori* classes, are detailed. We show the R packages used, data structures *distributionH* and *MatH* and the main functions developed. Moreover, both optimisation and multi-class classification technical specifications are explained.

- Chapter 5 focuses on assessing the models' performance in practice. We studied the application of the models developed on several data sets, either synthetic or real ones, both on histogram-valued and interval-valued data. Firstly, we give the description of the data and the context in which they are involved. The obtained results are presented and discussed.

- Finally, Chapter 6 summarises the problem at hand as well as the proposed extension for multi-class classification problems. The main conclusions of this work are presented, opening to future perspectives and opportunities for improvement.

Chapter 2

Symbolic Data Analysis

In this chapter, we start by introducing symbolic data, using classic data as the starting point, stating the main advantages of considering this data representation. Definitions and notation of symbolic variables are given, putting a special emphasis on histogram-valued variables, since those are the ones considered in this thesis. Furthermore, several concepts and functions used are presented and explained such as: quantile functions, specifying their purpose, arithmetic and assumptions required; the Mallows distance; the empirical symbolic mean; the choice of graphical representation of histogram-valued variables and the conclusions that may be drawn regarding the behaviour displayed. This is accompanied by examples using three data sets that aim at developing the intuitive point of view of the conclusions.

Moreover, the following section introduces useful descriptive measures. Well-known metrics such as the barycentric histogram, the inertia measure, empirical symbolic covariance, variance and linear correlation are presented.

New contributions are advanced in this chapter concerning the rewriting of the covariance and linear correlation *formulae* and the proof of three mathematical properties associated with the symbolic linear correlation measure.

The final section of this chapter clarifies the assumptions required for the method developed and reveals the definition of the linear combination of histogram-valued variables that leads the way to the linear discriminant function developed in the Chapter 3.

2.1 From Classic Data Analysis to Symbolic Data Analysis

In classical multivariate data analysis, data is represented by a $n \times p$ table with n individuals or observations, also called first-level units, and p variables or attributes. Each cell in the table is a concretisation of a variable that can be of a variety of types of data such as character, numeric or categorical. These data sets may be referenced as microdata.

In Symbolic Data Analysis (SDA), data are represented using symbolic variables. Each higher-level observation is called a unit and represents the concretisation of a symbolic variable. Data units no longer display only unique values, but have also internal variation and structure. These data sets may be referenced as macro data. Symbolic Data Analysis (SDA) aims to be a generalisation of classical/standard data analysis ([9], [10] and [11]).

Symbolic data are often obtained by aggregating microdata without applying a summary indicator. In such a manner, we hold on to the variability associated with the original data. At the same time, since we are dealing with more complex data, the current methods become no longer suitable. Therefore, it becomes emergent the development of new methods and models appropriate for the analysis of symbolic data.

Example 2.1. *Consider a data set where individuals are students (first-level units) with attributes such as age, gender, school and subjects' marks. If the statistical units of interest are the schools (higher-level units), an aggregation of data concerning students that attend the same school must be performed.*

2.2 Symbolic Variables: Definitions and Notation

We introduce the definition of symbolic variables presented in [9], [10] and [11].

Definition 2.1 (Symbolic Variable)**.** A symbolic variable X is a mapping

$$\begin{aligned} X : & A \to D \\ & i \to X(i) = h_i \end{aligned}$$

where A is a set of statistical entities. The set A may be of the form $A = \Omega = \{1, 2, \ldots, n\}$, in which case we have first-level units, or $A = \{C_1, C_2, \ldots, C_n\}$, with $C_i \subseteq \Omega$, in which case we have higher-level units (classes/concepts or categories). To each unit i in A there is a concretisation associated in D.

	X_1	X_2	...	X_j	...	X_p
1	h_{11}	h_{12}	...	h_{1j}	...	h_{1p}
2	h_{21}	h_{22}	...	h_{2j}	...	h_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
n	h_{n1}	h_{n2}	...	h_{nj}	...	h_{np}

TABLE 2.1: Symbolic data table.

Table 2.1 displays a complex data table used to represent symbolic data, considering p symbolic variables and n units. To clarify, the equivalent to the n individuals in classical data analysis are now designated as units.

X_j is used to reference the jth variable, for $j \in \{1, ..., p\}$. To reference the cell in this table that corresponds to the concretisation of the variable X_j on a unit i, with $i \in \{1, ..., n\}$, we used $X_j(i)$ or h_{ij}. When we have only one variable,X, this notation can be simplified. $X(i) = h_i$ is used to denote the value that the variable takes on the unit i, i.e., omit j.

Similarly to the classical data analysis, a symbolic description of unit i, with $i \in \{1, ..., n\}$, is given by the vector $(h_{i1}, h_{i2}, ..., h_{ij}, ..., h_{ip})$. Equivalently, it may be represented as $(X_1(i), X_2(i), ..., X_j(i), ..., X_p(i))$.

These variables are obtained from aggregation that can be of different types [12]:

- Temporal Aggregation: data are recorded at different moments in time for the same entities. However, suppose that we are not interested in the chronological order of observations. Consequently, we can aggregate the data without taking into account the temporal information. In this type of aggregation, the statistical units are the same before and after the aggregation.

- Contemporary Aggregation: data are recorded at the same point in time or the temporal instant is not documented. In this case, the aim of the study is to analyse entities at a higher level. In contemporary aggregation, the statistical units after the aggregation (higher-level units) differ from the initial ones (first-level units).

The aggregated individuals are, therefore, classes of units or observations that are mentioned as higher-level units.

Similarly to classical variables, symbolic variables can be either qualitative or quantitative. Let $\mathcal{Y}$ be the underlying set of X. According to the type of concretisation of the symbolic variable, the set D may be:

- $D = \mathcal{Y}$: single-valued variables - if each unit takes a single value. These variables correspond to the classical ones:
  - if $\mathcal{Y} \subseteq \mathbb{R}$, then we have a quantitative single-valued variable;
  - if $\mathcal{Y}$ is a set of categories, then we have a categorical single-valued variable;
- $D = \mathcal{P}(\mathcal{Y}) = \{E : E \subseteq \mathcal{Y}, E \neq \emptyset\}$, set of non-empty subsets of $\mathcal{Y}$: multi-valued variables - if each unit takes a finite set of values:
  - if $\mathcal{Y} \subseteq \mathbb{R}$ and the variable's concretisations are finite sets of real numbers, we call it a quantitative multi-valued variable;
  - if the variable's concretisations are finite sets of categories in $\mathcal{Y}$, then we have a categorical multi-valued variable;
- D is a set of intervals of values in $\mathcal{Y} \subseteq \mathbb{R}$: interval-valued variables - if each unit takes an interval of real values.
- D is a set of distributions of values on $\mathcal{Y}$: distributional-valued variables - if each unit displays a non-negative measure distribution. The measure (weight, relative frequency or probability) represents how frequent that category is for a given statistical entity:
  - if $\mathcal{Y} \subseteq \mathbb{R}$, then we have a quantitative distributional-valued variable;
  - $\mathcal{Y}$ is a set of categories and the values of each concretisation are in $\mathcal{Y}$, then we have a categorical distributional-valued variable;

Definition 2.2 (Histogram-Valued Variables)**.** The realisation of a histogram-valued variable X is a finite number of contiguous, disjoint and non-overlapping intervals that should be ordered. Each interval is associated with a non-negative weight that can be interpreted as a probability or frequency [11]. Recall that histograms represent empirical distributions where the values in each subinterval are assumed to be uniformly distributed.

Assuming that m is the number of subintervals, for each unit i, $X(i)$ is of the form:

$$X(i) = \{I_{X(i)1}, p_{X(i)1}; I_{X(i)2}, p_{X(i)2}; ...; I_{X(i)m}, p_{X(i)m}\},$$

where $I_{X(i)l}$ represents the subinterval l, $p_{X(i)l}$ is the non-negative weight associated with the subinterval $I_{X(i)l}$ and $\sum_{l=1}^{m} p_{X(i)l} = 1$.

Histogram-valued variables are a particular type of distributional-valued variables since each unit is represented by a finite number of subintervals, within each subinterval it is assumed a Uniform distribution and to each subinterval there is a probability associated, i.e., each unit displays an empirical distribution (histogram). Moreover, interval-valued variables are a particular type of histogram-valued variables. Consider a histogram-valued variable X with $m = 1$, i.e., only one subinterval. Since, $\sum_{l=1}^{m} p_{X(i)l} = 1$ and $m = 1$, then $p_{X(i)} = 1$. For each unit i, $X(i)$ is of the form:

$$X(i) = \{I_{X(i)}, 1\}$$

where $I_{X(i)}$ represents the interval that is the ith concretisation.

In other words, we obtain an interval-valued variable. In fact, interval-valued variables are a special case of histogram-valued variables.

The subinterval $I_{X(i)l}$ may be represented by its bounds or by its centre and (half)-range.

- Using interval bounds:

  $$I_{X(i)l} = [\underline{I}_{X(i)l}, \overline{I}_{X(i)l}[,$$

  where $\underline{I}_{X(i)l}$ is the lower bound and $\overline{I}_{X(i)l}$ is the upper bound of the subinterval l. For each subinterval l, $\underline{I}_{X(i)l} \leq \overline{I}_{X(i)l}$ for $l \in \{1, 2, ..., m\}$ and $\overline{I}_{X(i)l} \leq \underline{I}_{X(i)(l+1)}$ for $l \in \{1, 2, ..., m-1\}$.

- Using centres and half ranges:

  $$I_{X(i)l} = [c_{X(i)l} - r_{X(i)l}, c_{X(i)l} + r_{X(i)l}[,$$

  where $c_{X(i)l} = \frac{\overline{I}_{X(i)l} + \underline{I}_{X(i)l}}{2}$ is the centre and $r_{X(i)l} = \frac{\overline{I}_{X(i)l} - \underline{I}_{X(i)l}}{2}$ is the half range of the subinterval l.

The realisation of a histogram-valued variable can be represented by a traditional histogram or by the cumulative distribution function which is equivalent to the associated quantile function. From the uniformity hypothesis, it follows that these quantile functions are piecewise linear functions.

Remark 2.3. Interval-valued data are the most widely considered case and for which more models and methods have been explored.

Example 2.2. *Table 2.2 displays an example of symbolic data regarding schools. Although the gathered information may be related to individual students, the units of interest are the schools. In such a situation, the information associated with a specific school is obtained by aggregating data of students that attend the same school.*

School	*MathGrade*	*StudyHours perWeek*	*Gender*
1	{ [2,3[,0.3; [3,15[,0.7}	[1,10]	{F,0.15; M,0.85}
2	{[7,9[,0.4; [9,11],0.6 }	[1,5]	{F,0.5; M,0.5}
3	{[12,13[,0.2; [13,15[,0.8 }	[8,15]	{F,0.55; M,0.45}
4	{[17,18[,0.9; [18,20],0.1 }	[16,19]	{F,0.75; M,0.25}
5	{[5,6[,0.2; [6,7[,0.8}	[1,4]	{F,0.3; M,0.7}
6	{[10,12[,0.35; [12,12.5[,0.65}	[5,14]	{F,0.5; M,0.5}

TABLE 2.2: Example of symbolic data concerning schools.

The variable MathGrade is a histogram-valued variable that represents the distribution of grades in mathematics in each school. The variable StudyHoursperWeek is an interval-valued variable that represents the range of weekly hours spent studying mathematics. Finally, the variable Gender is a categorical modal variable that represents the gender proportion in each school.

2.3 Quantile Functions

As mentioned previously, the realisation of a histogram-valued variable can be represented by a traditional histogram. Nevertheless, this representation forces us to deal with histogram arithmetic which proved to be troublesome in [6]. Alternatively, the realisation of a histogram-valued variable can be represented by the quantile function, i.e., the inverse of the cumulative distribution function ([3], [4] and [7]).

Definition 2.4 (Quantile Function)**.** Let X be a random variable with cumulative distribution function F. The quantile function, Ψ, is defined in [13] by:

$$\Psi(q) = F^{-1}(q) = inf\{x : F(x) > q\}, q \in [0,1]$$

It is also called the inverse cumulative distribution function, F^{-1}.

Consider a histogram-valued variable, X. For each unit i, let $X(i)$ be the realisation of the mentioned variable, with m subintervals. This realisation may be represented by a cumulative distribution function, $F_{X(i)}(x)$. Assuming a Uniform distribution within subintervals, this function is given by:

$$F_{X(i)}(x) = \begin{cases} 0 & \text{if } x < \underline{I}_{X(i)1} \\ \frac{x-\underline{I}_{X(i)1}}{\overline{I}_{X(i)1}-\underline{I}_{X(i)1}} p_{X(i)1} & \text{if } \underline{I}_{X(i)1} \leq x < \overline{I}_{X(i)1} \\ p_{X(i)1} + \frac{x-\underline{I}_{X(i)2}}{\overline{I}_{X(i)2}-\underline{I}_{X(i)2}} p_{X(i)2} & \text{if } \underline{I}_{X(i)2} \leq x < \overline{I}_{X(i)2} \\ \vdots & \\ 1 & \text{if } x \geq \overline{I}_{X(i)m} \end{cases} \tag{2.1}$$

where $\sum_{l=1}^{m} p_{X(i)l} = 1$, $\underline{I}_{X(i)l}$ is the lower bound and $\overline{I}_{X(i)l}$ is the upper bound of the subinterval l. For each subinterval l, $\underline{I}_{X(i)l} \leq \overline{I}_{X(i)l}$ for $l \in \{1, 2, ..., m\}$ and $\overline{I}_{X(i)l} \leq \underline{I}_{X(i)(l+1)}$ for $l \in \{1, 2, ..., m-1\}$.

Assuming a Uniform distribution within subintervals, the quantile function may also be used to represent the realisation of this variable [7]. It is a piecewise linear function that can be represented as:

$$\Psi_{X(i)}(q) = \begin{cases} \underline{I}_{X(i)1} + \frac{q}{w_{X(i)1}} \left(\overline{I}_{X(i)1} - \underline{I}_{X(i)1} \right) & \text{if } 0 \leq q < w_{X(i)1} \\ \underline{I}_{X(i)2} + \frac{q-w_{X(i)1}}{w_{X(i)2}-w_{X(i)1}} \left(\overline{I}_{X(i)2} - \underline{I}_{X(i)2} \right) & \text{if } w_{X(i)1} \leq q < w_{X(i)2} \\ \vdots & \\ \underline{I}_{X(i)m} + \frac{q-w_{X(i)(m-1)}}{1-w_{X(i)(m-1)}} \left(\overline{I}_{X(i)m} - \underline{I}_{X(i)m} \right) & \text{if } w_{X(i)(m-1)} \leq q \leq 1 \end{cases} \tag{2.2}$$

or

$$\Psi_{X(i)}(q) = \begin{cases} c_{X(i)1} + r_{X(i)1} \left(\frac{2q}{w_{X(i)1}} - 1 \right) & \text{if } 0 \leq q < w_{X(i)1} \\ c_{X(i)2} + r_{X(i)2} \left(\frac{2(q-w_{X(i)1})}{w_{X(i)2}-w_{X(i)1}} - 1 \right) & \text{if } w_{X(i)1} \leq q < w_{X(i)2} \\ \vdots & \\ c_{X(i)m} + r_{X(i)m} \left(\frac{2(q-w_{X(i)(m-1)})}{1-w_{X(i)(m-1)}} - 1 \right) & \text{if } w_{X(i)(m-1)} \leq q \leq 1 \end{cases} \tag{2.3}$$

where $w_{X(i)l} = \sum_{d=1}^{l} p_{X(i)d}$ that are called cumulative weights.

Note that quantile functions are non-decreasing in their domain ($[0, 1]$). Moreover, the arithmetic associated with quantile functions is far less complex than histogram arithmetic.

Recall that interval-valued variables can be seen as a specific case of histogram-valued variables. Therefore, for interval-valued variables, the previously displayed quantile function can be simplified. Let $X(i)$ be the realisation of the interval-valued variable X. For each unit i, the quantile function associated with $X(i)$ is:

$$\Psi_{X(i)}(q) = \underline{I}_{X(i)} + q\left(\overline{I}_{X(i)} - \underline{I}_{X(i)}\right), q \in [0,1] \tag{2.4}$$

or

$$\Psi_{X(i)}(q) = c_{X(i)} + r_{X(i)}(2q-1), q \in [0,1] \tag{2.5}$$

We introduce the reasoning used to obtain the quantile function, assuming a Uniform distribution within the subintervals. Consider a subinterval l of the ith concretisation associated with the histogram-valued variable X, $I_{X(i)l} = [\underline{I}_{X(i)l}, \overline{I}_{X(i)l}[$. Let $p_{X(i)l}$ be the associated weight. Note that $p_{X(i)l} = w_{X(i)l} - w_{X(i)(l-1)}$. Assuming a Uniform distribution within $I_{X(i)l}$, then the cumulative distribution function is given by:

$$\begin{aligned}\underset{x \in I_{X(i)l}}{\forall} F_{X(i)}(x) &= \sum_{d=1}^{l-1} p_{X(i)d} + \frac{(x - \underline{I}_{X(i)l})p_{X(i)l}}{\overline{I}_{X(i)l} - \underline{I}_{X(i)l}} \\ &= w_{X(i)(l-1)} + \frac{(x - \underline{I}_{X(i)l})(w_{X(i)l} - w_{X(i)(l-1)})}{\overline{I}_{X(i)l} - \underline{I}_{X(i)l}}\end{aligned}$$

Taking into consideration that $c_{X(i)l} = \frac{\overline{I}_{X(i)l} + \underline{I}_{X(i)l}}{2}$, $r_{X(i)l} = \frac{\overline{I}_{X(i)l} - \underline{I}_{X(i)l}}{2}$ and $I_{X(i)l} = [\underline{I}_{X(i)l}, \overline{I}_{X(i)l}] = [c_{X(i)l} - r_{X(i)l}, c_{X(i)l} + r_{X(i)l}]$, the cumulative distribution function can be rewritten as:

$$\begin{aligned}\underset{x \in I_{X(i)l}}{\forall} F_{X(i)}(x) &= w_{X(i)(l-1)} + \frac{(x - c_{X(i)l} + r_{X(i)l})(w_{X(i)l} - w_{X(i)(l-1)})}{c_{X(i)l} + r_{X(i)l} - c_{X(i)l} + r_{X(i)l}} \\ &= w_{X(i)(l-1)} + \frac{(x - c_{X(i)l} + r_{X(i)l})(w_{X(i)l} - w_{X(i)(l-1)})}{2r_{X(i)l}}\end{aligned}$$

By developing the equation in order to x, within the mentioned interval, the quantile function, $\Psi_{X(i)}(q)$, is given by:

$$\Psi_{X(i)}(q) = F^{-1}_{X(i)}(q) = 2r_{X(i)l}\frac{q - w_{X(i)(l-1)}}{w_{X(i)l} - w_{X(i)(l-1)}} + c_{X(i)l} - r_{X(i)l}, q \in [w_{X(i)(l-1)}, w_{X(i)l}[$$

Therefore, we finally obtain the following formula:

$$\Psi_{X(i)}(q) = c_{X(i)l} + r_{X(i)l}\left(\frac{2(q - w_{X(i)(l-1)})}{w_{X(i)l} - w_{X(i)(l-1)}} - 1\right), q \in [w_{X(i)(l-1)}, w_{X(i)l}[$$

If any of the weights of $X(i)$, $p_{X(i)l}$, is null, then the quantile function $\Psi_{X(i)}(q)$ does not have inverse with domain between 0 and 1. In the scenario where we have one null weight, $\Psi_{X(i)}(q)$ is not continuous and has $m-1$ pieces.

The introduction of the quantile function that represents the symmetric histogram is crucial to grant an inverse linear relation in the linear combination of histogram-valued variables as it will be seen in due course.

Firstly, it stands to reason that two histograms $X(i)$ and $-X(i)$ are symmetric if they are geometrically symmetric with respect to the y-axis. Following this line of thought, the quantile function that represents the symmetric histogram (quantile function of $-X(i)$) is related to the quantile function of $X(i)$. Let $\Psi_{X(i)}(q)$ be the quantile function of the histogram-valued variable X, for $i \in \{1, ..., n\}$, then $-\Psi_{X(i)}(1-q)$ is the quantile function that represents its symmetric, $-X(i)$, as it is explored in [6]. Note that multiplying the quantile function $\Psi_{X(i)}(q)$ by -1 does not result in the quantile function that represents its symmetric. The function $-\Psi_{X(i)}(q)$ is not a non-decreasing function, therefore it is not a quantile function.

It is defined according to the expression 2.3 of the Definition 2.4 and it is given by:

$$-\Psi_{X(i)}(1-q) = \begin{cases} -c_{X(i)m} + r_{X(i)m}\left(\frac{2q}{w_{X(i)1}} - 1\right) & \text{if } 0 \leq q < w_{-X(i)1} \\ -c_{X(i)(m-1)} + r_{X(i)(m-1)}\left(\frac{2(q-w_{X(i)1})}{w_{X(i)2}-w_{X(i)1}} - 1\right) & \text{if } w_{-X(i)1} \leq q < w_{-X(i)2} \\ \vdots & \\ -c_{X(i)1} + r_{X(i)1}\left(\frac{2(q-w_{X(i)(m-1)})}{1-w_{X(i)(m-1)}} - 1\right) & \text{if } w_{-X(i)(m-1)} \leq q \leq 1 \end{cases} \tag{2.6}$$

Remark 2.5. When a given histogram is symmetric with respect to the y-axis, both the quantile function and quantile function that represents the symmetric histogram coincide, i.e., $\Psi_{X(i)}(q) = -\Psi_{X(i)}(1-q)$.

Example 2.3. *Consider the histogram-valued variable X and the first concretisation:*

$$X(1) = \{[2, 3[, 0.3; [3, 5[, 0.2; [5, 7[, 0.3; [7, 9[, 0.2\}$$

The quantile function associated with the first concretisation is:

$$\Psi_{X(1)}(q) = \begin{cases} 2.5 + 0.5\left(\frac{2q}{0.3} - 1\right) & \textit{if } 0 \leq q < 0.3 \\ 4 + 1\left(\frac{2(q-0.3)}{0.5-0.3} - 1\right) & \textit{if } 0.3 \leq q < 0.5 \\ 6 + 1\left(\frac{2(q-0.5)}{0.8-0.5} - 1\right) & \textit{if } 0.5 \leq q < 0.8 \\ 8 + 1\left(\frac{2(q-0.8)}{1-0.8} - 1\right) & \textit{if } 0.8 \leq q \leq 1 \end{cases}$$

Figure 2.1 displays the plot of the previous quantile function, the result of multiplying the quantile function by -1, referenced as Negative Quantile Function, and the quantile function that represents the histogram that is symmetric of the one represented by $\Psi_{X(1)}(q)$, referenced as Symmetric Quantile Function.

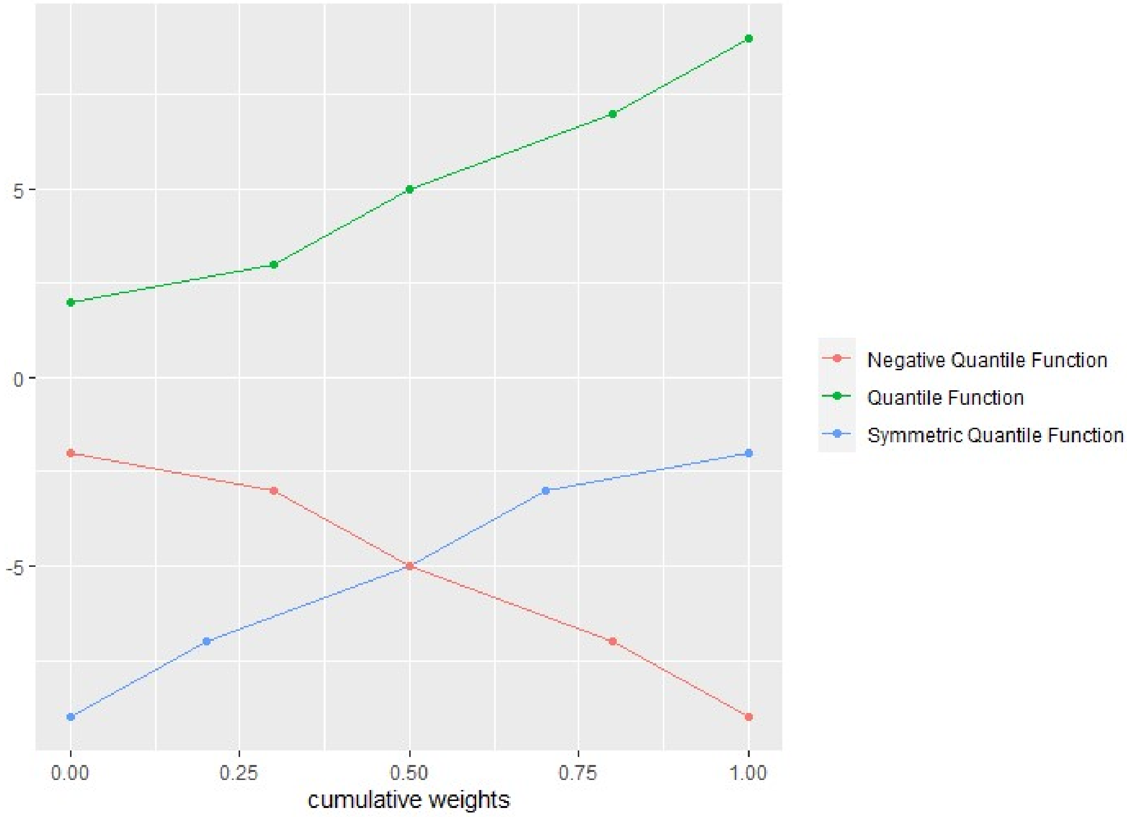


FIGURE 2.1: Representation of the functions $-\Psi_{X(1)}(q)$, $\Psi_{X(1)}(q)$ and $-\Psi_{X(1)}(1-q)$.

As it is possible to realise, the negative quantile function is a decreasing function, therefore it is not a quantile function. However, the symmetric quantile function is a non-decreasing function. This is an example that supports the definition of the symmetric quantile function of $\Psi_{X(i)}(q)$ as $-\Psi_{X(i)}(1-q)$ and not $-\Psi_{X(i)}(q)$.

2.4 Rewriting Operation

When using quantile functions as the representation of histograms, it is convenient to define all functions involved with an equal number of pieces and an equal domain for each piece. Therefore, all histograms must have an equal number of subintervals and equal weights associated with each corresponding subinterval. When this is not the case, a rewriting operation, introduced in [7], must be performed.

Consider the histogram-valued variable X. Let $X(i) = h_i$ denote the concretisation of the mentioned variable on unit i, for $i \in \{1, ..., n\}$. Moreover, let m_i be the number of subintervals of histogram h_i, with cumulative weights $\{w_{i0}, w_{i1}, \ldots, w_{im_i}\}$.

The first step in this rewriting operation is to identify the set of cumulative weights, Z, for all the units:

$$Z = \{w_{10}, w_{11}, \ldots, w_{1m_1}; w_{20}, w_{21}, \ldots, w_{2m_2}; \ldots; w_{n0}, w_{n1}, \ldots, w_{nm_n}\}$$

The second step is to sort Z without repetitions, leading to the set Z':

$$Z' = \{w_0, w_1, ..., w_l, ..., w_m\},$$

where $l \in \{0, ..., m\}$, $w_0 = 0$, $w_m = 1$ and $\max\{m_1, ..., m_n\} \leq m \leq \sum_{i=1}^{n} m_i - n + 1$.

The previous expression, regarding the worst possible case (when the weights are all different), considers the following reasoning: the set Z will have exactly $\sum_{i=1}^{n}(m_i + 1)$ cumulative weights. However, the weights $w_{i0} = 0$ and $w_{im_i} = 1$ are repeated $n - 1$ times. Since m is the number of subintervals that corresponds to the number of cumulative weights minus one, we obtain:

$$\begin{aligned} m &\leq \sum_{i=1}^{n}(m_i + 1) - 2(n-1) - 1 \\ \Leftrightarrow m &\leq \sum_{i=1}^{n} m_i + n - 2n + 2 - 1 \\ \Leftrightarrow m &\leq \sum_{i=1}^{n} m_i - n + 1 \end{aligned}$$

In such a manner, each histogram $X(i) = h_i$ can then be rewritten into one that, for subinterval l, has $\Psi_{X(i)}(w_{l-1})$ and $\Psi_{X(i)}(w_l)$ as lower and upper bounds, respectively.

In [7], it is displayed that the initial quantile function, $\Psi_{X(i)}(q)$, can also be written as a quantile function with m pieces:

$$\Psi_{X(i)}(q) = \begin{cases} \Psi_{X(i)}(0) + \frac{q}{w_1}\left(\Psi_{X(i)}(w_1) - \Psi_{X(i)}(0)\right) & \text{if } 0 \leq q < w_1 \\ \Psi_{X(i)}(w_1) + \frac{q-w_1}{w_2-w_1}\left(\Psi_{X(i)}(w_2) - \Psi_{X(i)}(w_1)\right) & \text{if } w_1 \leq q < w_2 \\ \vdots & \\ \Psi_{X(i)}(w_{m-1}) + \frac{q-w_{m-1}}{1-w_{m-1}}\left(\Psi_{X(i)}(1) - \Psi_{X(i)}(w_{m-1})\right) & \text{if } w_{m-1} \leq q \leq 1 \end{cases} \tag{2.7}$$

Remark 2.6. The rewriting operation must take into account, not only the concretisations of the symbolic variables but also the associated symmetric histograms since they are necessary for the linear discriminant method developed.

2.5 Operations with Quantile Functions

Apart from the fact that operations regarding quantile functions are simpler than those involving histograms, it is also easier to operate with quantile functions that have an equal number of pieces and an equal domain for each piece. We introduce several operations regarding quantile functions, displayed in [6].

Consider two histogram-valued variables X and Y. For a given unit i, let $X(i)$ be the realisation of the variable X and, for a given unit j, let $Y(j)$ be the realisation of the variable Y, with associated quantile functions $\Psi_{X(i)}$ and $\Psi_{Y(j)}$, respectively. Moreover, assume that the quantile functions have an equal number of pieces, m, and an equal domain for each piece. The ordered set of cumulative weights can be written as $\{0, w_1, ..., w_{m-1}, 1\}$.

2.5.1 Operation of Addition

The operation of addition between two quantile functions is defined as:

$$\Psi_{X(i)}(q) + \Psi_{Y(j)}(q) = \begin{cases} c_{X(i)1} + c_{Y(j)1} + (r_{X(i)1} + r_{Y(j)1})\left(\frac{2q}{w_1} - 1\right) & \text{if } 0 \leq q < w_1 \\ c_{X(i)2} + c_{Y(j)2} + (r_{X(i)2} + r_{Y(j)2})\left(\frac{2(q-w_1)}{w_2-w_1} - 1\right) & \text{if } w_1 \leq q < w_2 \\ \vdots & \\ c_{X(i)m} + c_{Y(j)m} + (r_{X(i)m} + r_{Y(j)m})\left(\frac{2(q-w_{(m-1)})}{1-w_{(m-1)}} - 1\right) & \text{if } w_{(m-1)} \leq q \leq 1 \end{cases} \tag{2.8}$$

In this case, we always obtain a non-decreasing function. Both the slopes and the y-intercepts of the resulting quantile function are affected by both initial quantile functions [6].

The operation of addition between a quantile function and a real value α is defined as:

$$\Psi_{X(i)}(q)+\alpha=\begin{cases} c_{X(i)1}+\alpha+r_{X(i)1}\left(\frac{2q}{w_1}-1\right) & \text{if } 0\leq q<w_1 \\ c_{X(i)2}+\alpha+r_{X(i)2}\left(\frac{2(q-w_1)}{w_2-w_1}-1\right) & \text{if } w_1\leq q<w_2 \\ \vdots & \\ c_{X(i)m}+\alpha+r_{X(i)m}\left(\frac{2(q-w_{(m-1)})}{1-w_{(m-1)}}-1\right) & \text{if } w_{(m-1)}\leq q\leq 1 \end{cases} \tag{2.9}$$

This operation corresponds to a translation of the histogram $X(i)$ in a direction parallel to the horizontal axis with a distance provided by the value α [6].

2.5.2 Operation of Multiplication

The operation of multiplication of a quantile function by a real value α is defined as:

$$\alpha\Psi_{X(i)}(q)=\begin{cases} \alpha c_{X(i)1}+\alpha r_{X(i)1}\left(\frac{2q}{w_1}-1\right) & \text{if } 0\leq q<w_1 \\ \alpha c_{X(i)2}+\alpha r_{X(i)2}\left(\frac{2(q-w_1)}{w_2-w_1}-1\right) & \text{if } w_1\leq q<w_2 \\ \vdots & \\ \alpha c_{X(i)m}+\alpha r_{X(i)m}\left(\frac{2(q-w_{(m-1)})}{1-w_{(m-1)}}-1\right) & \text{if } w_{(m-1)}\leq q\leq 1 \end{cases} \tag{2.10}$$

In this scenario, both the slopes and the y-intercepts are affected by α. Additionally, if $\alpha>0$, then we obtain a non-decreasing function. In the case where $\alpha<0$, we obtain a decreasing function. For this reason, the result of an operation of multiplication by a real value may not generate a quantile function [6].

2.5.3 Space of Quantile Functions

The space of quantile functions is a semi-vector space [14].

Definition 2.7. A semi-vector space (over $\mathbb{R}^+$) is defined to be a set $\mathcal{U}$ equipped with the operations $+:\mathcal{U}\times\mathcal{U}\longrightarrow\mathcal{U}$ and $\cdot:\mathbb{R}^+\times\mathcal{U}\longrightarrow\mathcal{U}$ such that the following properties are satisfied. For each $r,s,\in\mathbb{R}^+$, $u,v,w\in\mathcal{U}$,

$$u + (v + w) = (u + v) + w, \qquad u + v = v + u,$$

$$(rs)u = r(su) \qquad 1u = u$$

$$r(u + v) = ru + rv \qquad (r + s)u = ru + su$$

2.6 Mallows Distance

In order to define the linear discriminant functions an optimisation step is required. The optimised parameters are determined considering a criterion based on the Mallows Distance, also known as Earth Mover's distance, introduced in [15]. This distance is considered to be a good measure for evaluating the similarity between distributions. It is used since the calculation is simple ([4], [7] and [12]).

Definition 2.8. Consider X and Y, histogram-valued variables. For a given unit i, let $X(i)$ be the realisation of the variable X and, for a given unit j, let $Y(j)$ be the realisation of the variable Y, represented by the quantile functions $\Psi_{X(i)}$ and $\Psi_{Y(j)}$, both with m pieces and the same set of weights, $\{p_1, \ldots, p_m\}$. The Mallows distance presented in [7] and [12] is defined as follows:

$$D_M(\Psi_{X(i)}, \Psi_{Y(j)}) = \sqrt{\int_0^1 (\Psi_{X(i)}(q) - \Psi_{Y(j)}(q))^2 \, dq} \tag{2.11}$$

The Mallows Distance is an adequate measure since it has an intuitive interpretation, adjusting to the concept of distance assessed by the human eye [16].

Assuming a Uniform distribution within subintervals, it can be proven that the expression 2.11 can be rewritten as:

$$D_M(\Psi_{X(i)}, \Psi_{Y(j)}) = \sqrt{\sum_{l=1}^{m} p_l \left[(c_{X(i)l} - c_{Y(j)l})^2 + \frac{1}{3}(r_{X(i)l} - r_{Y(j)l})^2 \right]} \tag{2.12}$$

The detailed proof of the simplification of the formula for the Mallows distance, expression 2.12, can be found in [6].

In the case of interval-valued variables, the number of subintervals, m, is 1 with associated weight $p_l = 1$, hence the expression 2.12 can be simplified:

$$D_M(\Psi_{X(i)}, \Psi_{Y(j)}) = \sqrt{(c_{X(i)} - c_{Y(j)})^2 + \frac{1}{3}(r_{X(i)} - r_{Y(j)})^2} \tag{2.13}$$

2.7 Graphical Display of Symbolic Variables

Thus far, only analytical aspects of Symbolic Data Analysis have been addressed. Nevertheless, two histogram-valued variables may be displayed in a scatter plot as it is explored in [6]. A graphical display of histogram-valued variables is important to visually reveal the relation between them.

The representation of two histogram-valued variables, X and Y, may be done using a scatter plot. Each unit is described by a concretisation of both variables X and Y. This pair of concretisations can be plotted in a 2D figure. Assuming that the histograms that represent $X(i)$ and $Y(i)$, for each unit $i \in \{1, \ldots, n\}$, have the same number of subintervals, each subinterval is represented by a rectangle. Particularly, the scatter plot of two histogram-valued variables that only have one concretisation corresponds graphically to a set of non-overlapping and contiguous rectangles.

From the scatter plot, it is possible to draw conclusions on the behaviour displayed. The concretisations for a pair of variables may exhibit a linear behaviour if the histograms (or their empirical symbolic mean values) are aligned in the scatter plot.

Definition 2.9 (Empirical Symbolic Mean)**.** Consider the histogram-valued variable X with n units and m subintervals. The empirical symbolic mean for X, introduced in [17], is given by:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{m} \frac{\overline{I}_{X(i)l} + \underline{I}_{X(i)l}}{2} p_{X(i)l} \tag{2.14}$$

For a specific unit i, for $i \in \{1, \ldots, n\}$, the empirical symbolic mean for $X(i)$ is given by:

$$\overline{X(i)} = \sum_{l=1}^{m} \frac{\overline{I}_{X(i)l} + \underline{I}_{X(i)l}}{2} p_{X(i)l} \tag{2.15}$$

The direction of the alignment determines if the relation between the two variables is direct or inverse. Assuming that X and Y exhibit a linear behaviour, the histogram-valued variable X is said to have a direct relation on the histogram-valued variable Y if the slope of the curves of empirical symbolic mean values of the concretisations of the mentioned variables is always positive. Otherwise, X is said to have an inverse relation on Y [6].

Example 2.4. *The purpose of this example is to show variables that exhibit a positive linear relation (direct), variables that exhibit a negative linear relation (inverse) and variables that do not exhibit either positive or negative linear relation.*

Consider three different symbolic data sets. Each symbolic data set developed has five units described by two histogram-valued variables. Consider the symbolic data Tables 2.3, 2.4 and 2.5 and the produced scatter plots 2.2, 2.3 and 2.4. Figures 2.2, 2.3 and 2.4 display the plots of the data associated with the Tables 2.3, 2.4 and 2.5, respectively, distinguishing each unit with a distinctive colour. In addition, we indicated, in black, the curve linking the sample mean values of the concretisations (histograms).

	X	Y
1	{ [2,3[,0.2; [3,5[,0.47; [5,7[,0.03; [7,9[,0.1; [9,11],0.2 }	{ [-2,-1[,0.2; [-1,0[,0.47; [0,1[,0.03;[1,2[,0.1; [2,4],0.2 }
2	{ [12,13[,0.2; [13,15[,0.47; [15,17[,0.03; [17,18[,0.1; [18,21],0.2 }	{ [19,22[,0.2; [22,24[,0.47; [24,26[,0.03;[26,28[,0.1; [28,33],0.2 }
3	{ [5,6[,0.2; [6,7[,0.47; [7,8[,0.03;[8,9[,0.1; [9,11],0.2 }	{ [-5,-1[,0.2; [-1,0[,0.47; [0,4[,0.03;[4,7[,0.1; [7,10],0.2 }
4	{ [7,9[,0.2; [9,11[,0.47; [11,13[,0.03; [13,14[,0.1; [14,15],0.2 }	{ [10,11[,0.2; [11,15[,0.47; [15,16[,0.03;[16,18[,0.1; [18,20],0.2 }
5	{ [10,12[,0.2; [12,12.5[,0.47; [12.5,15[,0.03;[15,16.5[,0.1; [16.5,17],0.2 }	{ [14,15[,0.2; [15,17[,0.47; [17,20[,0.03;[20,23[,0.1; [23,25],0.2 }

TABLE 2.3: Symbolic data table with direct relation.

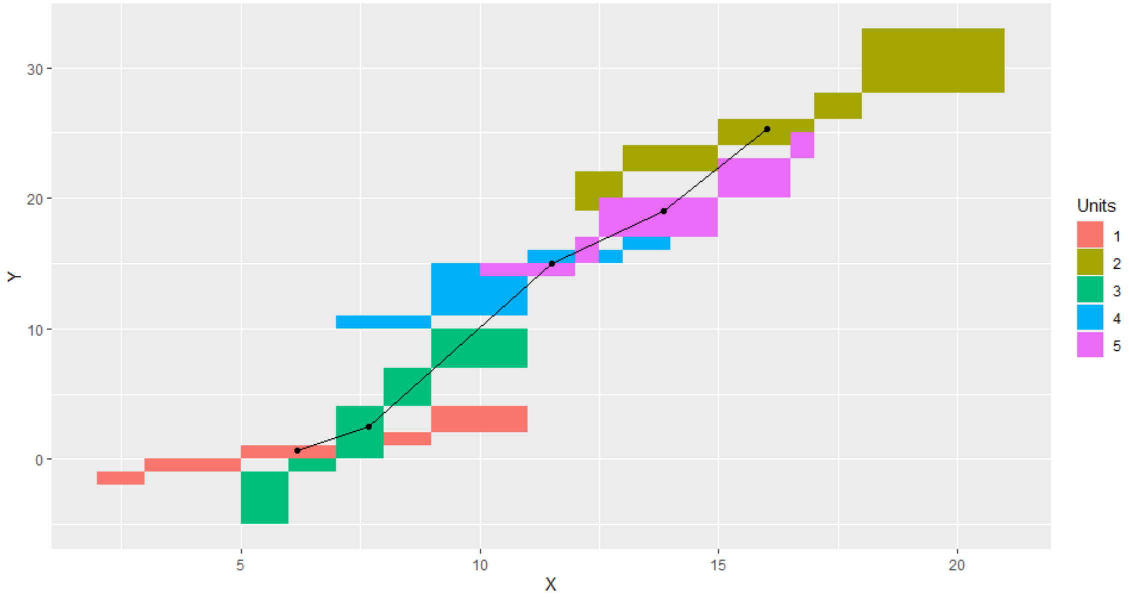


FIGURE 2.2: Scatter plot of the variables associated with the symbolic data table with direct relation.

	X	Y
1	{ [2,3[,0.2; [3,5[,0.47; [5,7[,0.03; [7,9[,0.1; [9,11],0.2 }	{ [2,3[,0.2; [3,5[,0.47; [5,7[,0.03; [7,9[,0.1; [9,11],0.2 }
2	{ [12,13[,0.2; [13,15[,0.47; [15,17[,0.03; [17,18[,0.1; [18,21],0.2 }	{ [-33,-28[,0.2; [-28,-26[,0.47; [-26,-24[,0.03; [-24,-22[,0.1; [-22,-19],0.2 }
3	{ [5,6[,0.2; [6,7[,0.47; [7,8[,0.03;[8,9[,0.1; [9,11],0.2 }	{ [-10,-7[,0.2; [-7,-4[,0.47; [-4,0[,0.03;[0,1[,0.1; [1,5],0.2 }
4	{ [7,9[,0.2; [9,11[,0.47; [11,13[,0.03; [13,14[,0.1; [14,15],0.2 }	{ [-20,-18[,0.2; [-18,-16[,0.47; [-16,-15[,0.03; [-15,-11[,0.1; [-11,-10],0.2 }
5	{ [10,12[,0.2; [12,12.5[,0.47; [12.5,15[,0.03; [15,16.5[,0.1; [16.5,17],0.2 }	{ [-25,-23[,0.2; [-23,-20[,0.47; [-20,-17[,0.03; [-17,-15[,0.1; [-15,-14],0.2 }

Table 2.4: Symbolic data table with inverse relation.

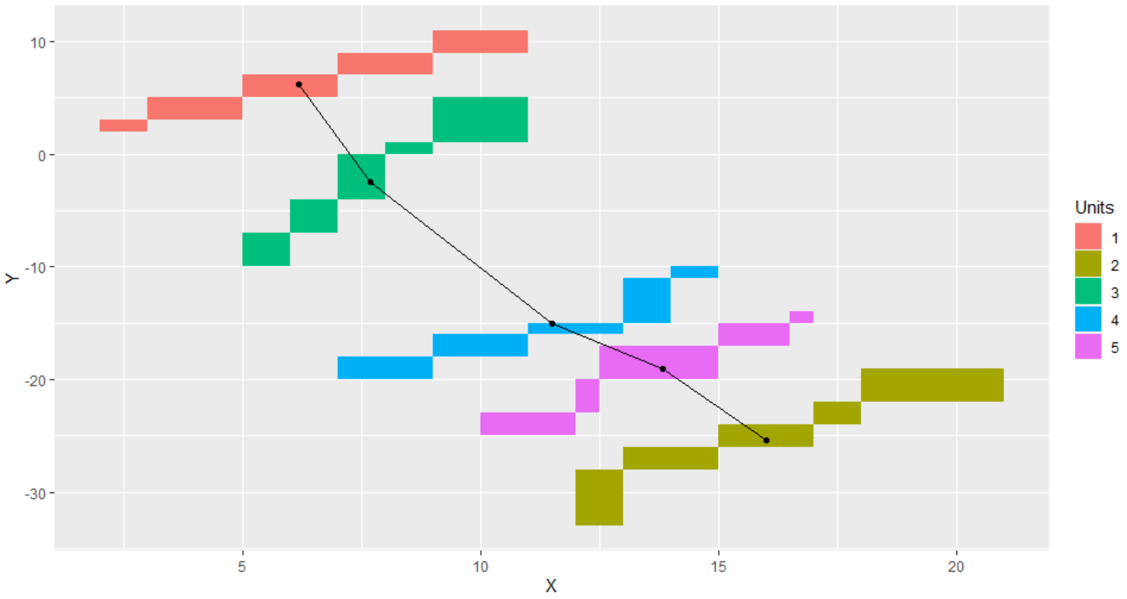


Figure 2.3: Scatter plot of the variables associated with the symbolic data table with inverse relation.

	X	Y
1	{ [-12,-10[,0.2; [-10,-7[,0.47; [-7,7[,0.03; [7,9[,0.1; [9,11],0.2 }	{ [12,13[,0.2; [13,15[,0.47; [15,17[,0.03; [17,19[,0.1; [19,21],0.2 }
2	{ [-33,-28[,0.2; [-28,-26[,0.47; [-26,-24[,0.03; [-24,-22[,0.1; [-22,-19],0.2 }	{ [-33,-28[,0.2; [-28,-26[,0.47; [-26,-24[,0.03; [-24,-22[,0.1; [-22,-19],0.2 }
3	{ [-10,-7[,0.2; [-7,-4[,0.47; [-4,0[,0.03;[0,1[,0.1; [1,5],0.2 }	{ [-10,-7[,0.2; [-7,-4[,0.47; [-4,0[,0.03;[0,1[,0.1; [1,5],0.2 }
4	{ [-20,-18[,0.2; [-18,-16[,0.47; [-16,-15[,0.03; [-15,-11[,0.1; [-11,-10],0.2 }	{ [-20,-18[,0.2; [-18,-16[,0.47; [-16,-15[,0.03; [-15,-11[,0.1; [-11,-10],0.2 }
5	{ [-25,-23[,0.2; [-23,-20[,0.47; [-20,-17[,0.03; [-17,-15[,0.1; [-15,-14],0.2 }	{ [-25,-23[,0.2; [-23,-20[,0.47; [-20,-17[,0.03; [-17,-15[,0.1; [-15,-14],0.2 }

Table 2.5: Symbolic data table without clear direct or inverse relation.

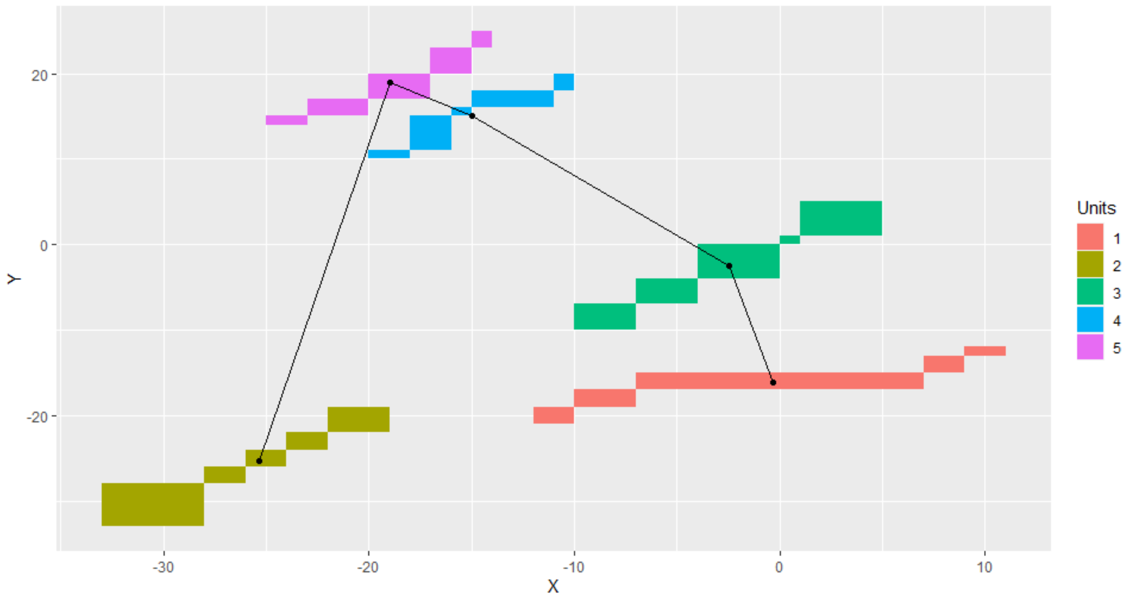


FIGURE 2.4: Scatter plot of the variables associated with the symbolic data table without clear direct or inverse relation.

In Figure 2.2, although the slope of the curve is not constant, it is consistently positive. Therefore, X and Y show a direct linear relation. Similarly, in Figure 2.3, the slope of the curve is consistently negative. Consequently, the variables involved show an inverse linear relation. On the other hand, in Figure 2.4, the slope of the curve alternates in sign. This irregularity reveals a lack of linear relation.

The linear relation described in this section is properly measured by the empirical symbolic linear correlation coefficient that will be explored in a following section.

2.8 Descriptive Measures

Descriptive measures are crucial to provide general characteristics that symbolic data may exhibit. In order to provide an efficient and meaningful way to describe and summarise the symbolic data, measures such as inertia measure and symbolic covariance were developed.

2.8.1 Barycentric Histogram

In descriptive statistics, the mean can take several forms. Since we are mainly dealing with histogram-valued variables, it stands to reason that the notion of mean should assume the form of a histogram. Considering the proximity among data, we can operate with the barycentric histogram. This is important in Linear Discriminant Analysis, because, in the method developed in [1], the classification of each unit is based on the distance between the unit's score to the barycentric histogram of each *a priori* class's score.

Definition 2.10 (Barycentric Histogram)**.** Consider n histograms of the histogram-valued variable X, $\{h_1, h_2, ..., h_n\}$, that follow a Uniform distribution within subintervals, the quantile functions have m pieces and the same set of cumulative weights, $\{0, w_1, w_2, \ldots, w_{(m-1)}, 1\}$. The quantile function of the barycentric histogram (also called Global Barycentre or simply Barycentre), $\overline{\Psi_X}$, is defined as:

$$\overline{\Psi_X}(q) = \underset{\Psi_X(q)}{\arg\min} \sum_{k=1}^{n} D_M^2(\Psi_{h_k}(q), \Psi_X(q)) \tag{2.16}$$

In other words, the Barycentric Histogram, proposed by [7], is the one that is at a minimum distance from all the others. This leads to the notion of "centre of gravity" of the set of histograms considered (centroid).

The optimal solution is obtained by solving a Least Squares Problem, resulting in the following formula:

$$\overline{\Psi_X}(q) = \begin{cases} \bar{c}_{X1} + \bar{r}_{X1}\left(\frac{2q}{w_1} - 1\right) & \text{if } 0 \leq q < w_1 \\ \bar{c}_{X2} + \bar{r}_{X2}\left(\frac{2(q-w_1)}{w_2-w_1} - 1\right) & \text{if } w_1 \leq q < w_2 \\ \vdots & \\ \bar{c}_{Xm} + \bar{r}_{Xm}\left(\frac{2(q-w_{(m-1)})}{1-w_{(m-1)}} - 1\right) & \text{if } w_{(m-1)} \leq q \leq 1 \end{cases} \tag{2.17}$$

where $\bar{c}_{Xl} = \frac{1}{n}\sum_{k=1}^{n} c_{h_k l}$ and $\bar{r}_{Xl} = \frac{1}{n}\sum_{k=1}^{n} r_{h_k l}$. In other words, these make use of the classical mean *formulae*.

Example 2.5. *Consider the symbolic data Table 2.4, with five units described by two histogram-valued variables.*

Figures 2.5, 2.6 and 2.7 display the barycentric histograms in pink. The plots offer visual evidence by way of explanation for the notion of centroid, regarding this descriptive measure.

2.8.2 Inertia Measure

While the Barycentric Histogram conveys the notion of central tendency, the Total Inertia is a measure of dispersion, introduced in [7].

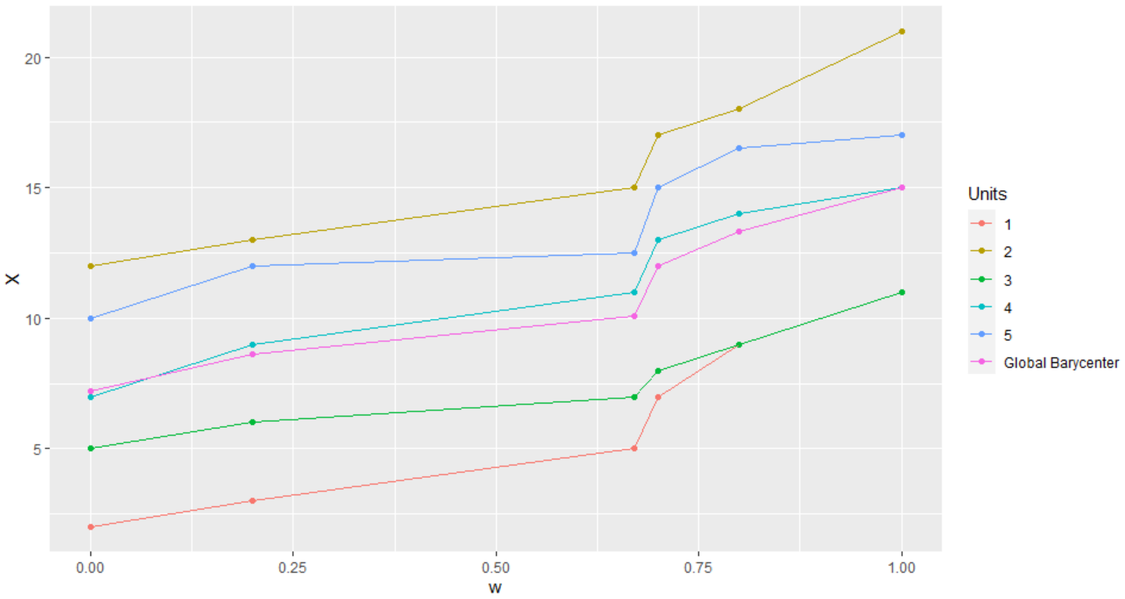


FIGURE 2.5: Quantile functions of the variable X and respective barycentric histogram for the data in Table 2.4.

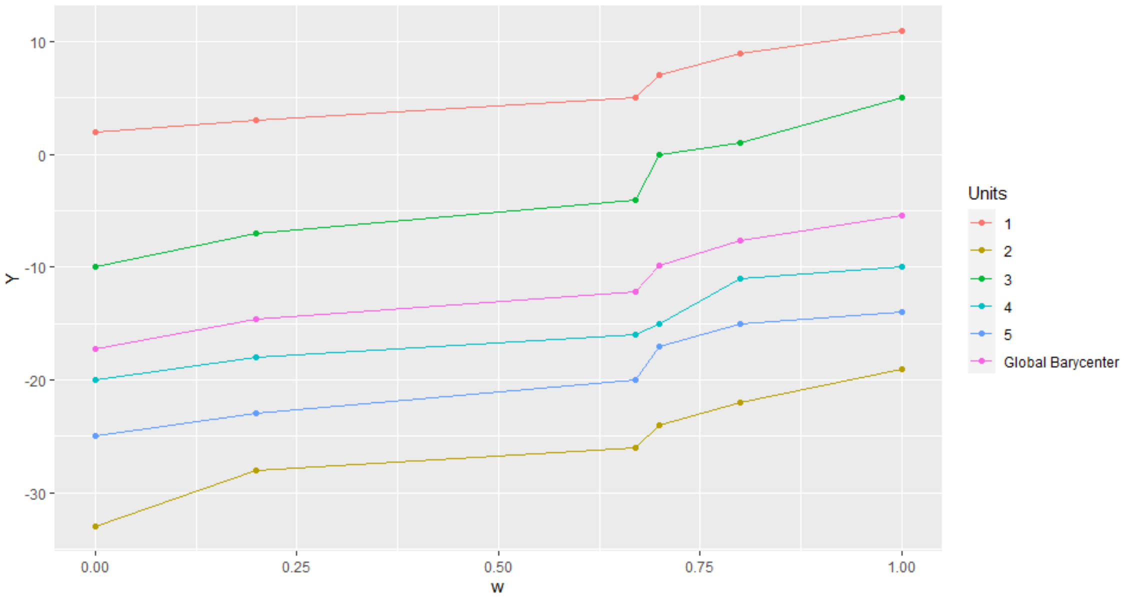


FIGURE 2.6: Quantile functions of the variable Y and respective barycentric histogram for the data in Table 2.4.

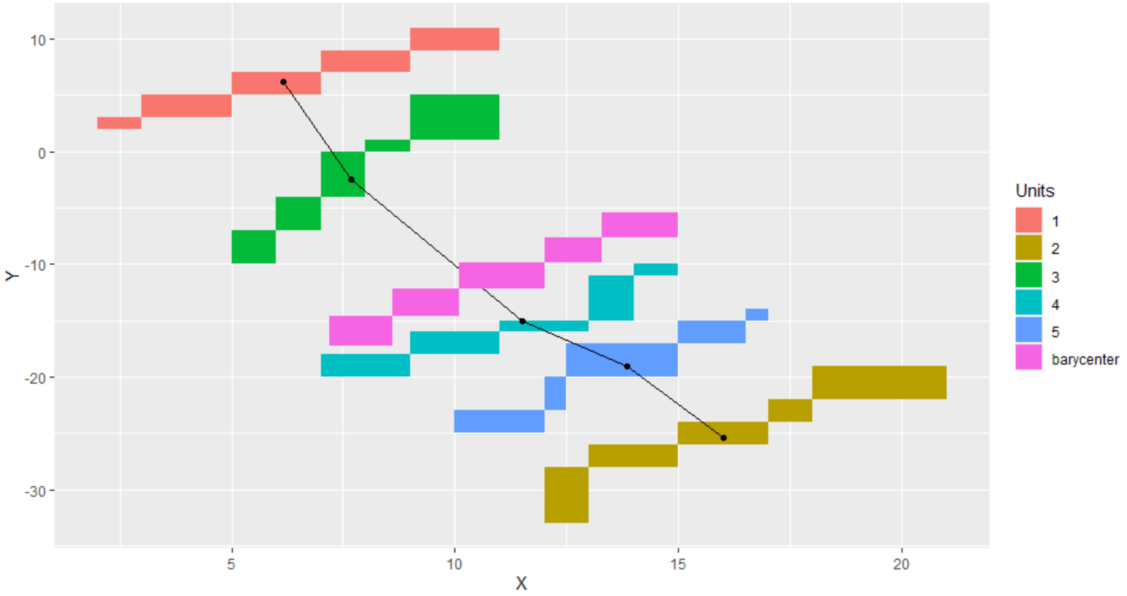


FIGURE 2.7: Scatter plot of the variables in Table 2.4 and respective barycentric histograms.

Definition 2.11 (Total Inertia)**.** The Total Inertia, TI, with respect to the Barycentric Histogram, $\overline{\Psi_X}$, associated with the histogram-valued variable X of a set of n histogram observations $\{h_1, h_2, ..., h_n\}$ is given by:

$$TI = \sum_{k=1}^{n} D_M^2(\Psi_{h_k}, \overline{\Psi_X}) \tag{2.18}$$

The choice of the Mallows distance empowers this dispersion measure through the Huygens theorem of decomposition of clustered histogram-valued data [7]. This theorem provides a basis to decompose the Total Inertia into Between Inertia, BI and Within Inertia, WI. Consider histogram-valued data with s clusters:

$$\begin{aligned} TI &= BI + WI \\ &= \sum_{u=1}^{s} n_u D_M^2(\overline{\Psi_{X_u}}, \overline{\Psi_X}) + \sum_{u=1}^{s} \sum_{h_k \in G_u} D_M^2(\Psi_{h_k}, \overline{\Psi_{X_u}}), \end{aligned} \tag{2.19}$$

where $u \in \{1, ..., s\}$, n_u is the cardinality of the group u, $\overline{\Psi_X}$ is the quantile function associated with the global barycentric histogram and $\overline{\Psi_{X_u}}$ is the quantile function associated with the barycentric histogram of the group of units that have *a priori* class u [7].

2.8.3 Symbolic Covariance

In addition to the previous dispersion measure, we introduce the symbolic covariance, defined in [8]. This is just a step towards the goal of defining the symbolic linear correlation coefficient that is used in the optimisation step.

Definition 2.12 (Symbolic Covariance)**.** Consider two random histogram-valued variables X and Y with n units and quantile functions $\Psi_{X(i)}$ and $\Psi_{Y(i)}$, for $i \in \{1, \ldots, n\}$. Moreover, let $\overline{\Psi_X}$ and $\overline{\Psi_Y}$ be the quantile function associated with the global barycentric histogram of the histogram-valued variables X and Y, respectively. The empirical covariance between these variables may be defined as follows:

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 (\Psi_{X(i)}(q) - \overline{\Psi_X}(q))(\Psi_{Y(i)}(q) - \overline{\Psi_Y}(q))\, dq \tag{2.20}$$

Proposition 2.13. *Consider two random histogram-valued variables X and Y with n units and quantile functions $\Psi_{X(i)}$ and $\Psi_{Y(i)}$, for $i \in \{1, \ldots, n\}$. Assume that both are written with m pieces, the same set of weights,$\{p_1, p_2, ..., p_m\}$, and a Uniform distribution within each subinterval. Given the paired concretisations $(X(1), Y(1)), \ldots, (X(n), Y(n))$, that is, n units, the symbolic covariance formula can be rewritten as:*

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl}) \right] \tag{2.21}$$

where $c_{X(i)l}$ and $c_{Y(i)l}$, $r_{X(i)l}$ and $r_{Y(i)l}$ are the centres and half ranges of the subinterval l associated with the ith concretisation of the histogram-valued variable X and Y, respectively. $\bar{c}_{Xl}$ and $\bar{c}_{Yl}$, $\bar{r}_{Xl}$ and $\bar{r}_{Yl}$ are the centres and half ranges of the subinterval l associated with the Barycentric Histogram of X and Y, respectively.

Proof. Consider a fixed $i \in \{1, \ldots, n\}$. According to the Definition 2.12, we have:

$$\int_0^1 (\Psi_{X(i)}(q) - \overline{\Psi_X}(q))(\Psi_{Y(i)}(q) - \overline{\Psi_Y}(q))\, dq \tag{2.22}$$

Consider the variable X. For each subinterval l and $q \in [w_{(l-1)}, w_l[$, $\Psi_{X(i)}(q) - \overline{\Psi_X}(q)$ can be developed into:

$$\begin{aligned}&\left(c_{X(i)l} + \left(\frac{2(q-w_{l-1})}{w_l - w_{l-1}} - 1\right) r_{X(i)l}\right) - \left(\bar{c}_{Xl} + \left(\frac{2(q-w_{l-1})}{w_l - w_{l-1}} - 1\right) \bar{r}_{Xl}\right)\\ &= \left((c_{X(i)l} - \bar{c}_{Xl}) + \left(\frac{2(q-w_{l-1})}{w_l - w_{l-1}} - 1\right)(r_{X(i)l} - \bar{r}_{Xl})\right)\end{aligned} \tag{2.23}$$

Similarly, this can be performed for the variable Y. Therefore, expression 2.22 is equal to:

$$\begin{aligned}\sum_{l=1}^{m} \int_{w_{(l-1)}}^{w_l} &\left(c_{X(i)l} - \bar{c}_{Xl} + \left(\frac{2(q-w_{l-1})}{w_l - w_{l-1}} - 1\right)(r_{X(i)l} - \bar{r}_{Xl})\right) \times \\ &\times \left(c_{Y(i)l} - \bar{c}_{Yl} + \left(\frac{2(q-w_{l-1})}{w_l - w_{l-1}} - 1\right)(r_{Y(i)l} - \bar{r}_{Yl})\right) dq\end{aligned} \tag{2.24}$$

Moreover, consider the change of variable:

$$v = \frac{q - w_{l-1}}{w_l - w_{l-1}}$$

$$\begin{aligned} dv &= \frac{1}{w_l - w_{(l-1)}}\, dq \\ &\Leftrightarrow (w_l - w_{(l-1)})\, dv = dq \\ &\Leftrightarrow p_l\, dv = dq \end{aligned}$$

Note that if $q = w_{(l-1)}$, then $v = 0$ and if $q = w_l$, then $v = 1$. The expression 2.24 can be rewritten as:

$$\begin{aligned}&\sum_{l=1}^{m} p_l \int_0^1 \left((c_{X(i)l} - \bar{c}_{Xl}) + (2v-1)(r_{X(i)l} - \bar{r}_{Xl})\right)\left((c_{Y(i)l} - \bar{c}_{Yl}) + (2v-1)(r_{Y(i)l} - \bar{r}_{Yl})\right) dv \\ &= \sum_{l=1}^{m} p_l \int_0^1 \left((c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl})\right) + \left((c_{X(i)l} - \bar{c}_{Xl})(2v-1)(r_{Y(i)l} - \bar{r}_{Yl})\right) + \\ &\quad + \left((c_{Y(i)l} - \bar{c}_{Yl})(2v-1)(r_{X(i)l} - \bar{r}_{Xl})\right) + \left((2v-1)^2 (r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl})\right) dv \\ &= \sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl})\right]\end{aligned}$$

Since we have n units, from the above equation, we can rewrite the definition of covariance as:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl}) \right]$$

□

Considering this definition of symbolic covariance, it is also possible to define the symbolic variance (and standard deviation) of a histogram-valued variable.

Definition 2.14 (Symbolic Variance)**.** Consider a random histogram-valued variable X, with n units, quantile functions $\Psi_{X(i)}$, for $i \in \{1, \ldots, n\}$, and the quantile function associated with the global barycentric histogram, $\overline{\Psi_X}$. Assuming a Uniform distribution within subintervals, the symbolic variance of X may be defined as follows:

$$\begin{aligned} var(X) &= cov(X, X) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (\Psi_{X(i)}(q) - \overline{\Psi_X}(q))(\Psi_{X(i)}(q) - \overline{\Psi_X}(q))\, dq \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (\Psi_{X(i)}(q) - \overline{\Psi_X}(q))^2\, dq \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2 \right] \\ &= \frac{1}{n}\sum_{i=1}^{n} D_M^2(\Psi_{X(i)}, \overline{\Psi_X}) \end{aligned} \tag{2.25}$$

Definition 2.15 (Symbolic Standard Deviation)**.** Consider a random histogram-valued variable X. Considering the Definition 2.14 and assuming a Uniform distribution within subintervals, the symbolic standard deviation of X may be defined as:

$$s_X = \sqrt{var(X)} = \sqrt{cov(X, X)} \tag{2.26}$$

2.8.4 Symbolic Linear Correlation

One of the approaches to extend the classification to more than two classes relies on the identification of several discriminant functions - Consecutive Linear Discriminant Functions. These functions are required to be uncorrelated.

2.8.4.1 Definition

The development of a linear correlation coefficient was crucial to measure how much two histogram-valued variables are linearly correlated. The symbolic linear correlation coefficient developed is similar to the correlation index developed in [8].

Definition 2.16 (Symbolic Correlation Coefficient)**.** Consider two random histogram valued variables X and Y. The symbolic correlation between these variables may be defined as follows:

$$r'(X,Y) = \frac{cov(X,Y)}{s_X s_Y}, \tag{2.27}$$

where:

- $cov(X,Y)$ stands for the symbolic covariance between the histogram-valued variables X and Y.
- s_X stands for the symbolic standard deviation of the histogram-valued variable X.
- s_Y stands for the symbolic standard deviation of the histogram-valued variable Y.

Proposition 2.17. *Let X and Y be two random histogram-valued variables with n units and quantile functions $\Psi_{X(i)}$ and $\Psi_{Y(i)}$, for $i \in \{1, \ldots, n\}$. Assume that both are written with m pieces, the same set of weights, $\{p_1, p_2, ..., p_m\}$, and a Uniform distribution within each subinterval. Considering the expressions 2.21 and 2.26, the symbolic correlation coefficient formula can be written as:*

$$\frac{\frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl})\right]}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2\right]}\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l \left[(c_{Y(i)l} - \bar{c}_{Yl})^2 + \frac{1}{3}(r_{Y(i)l} - \bar{r}_{Yl})^2\right]}} \tag{2.28}$$

where $c_{X(i)l}$ and $c_{Y(i)l}$, $r_{X(i)l}$ and $r_{Y(i)l}$ are the centres and half ranges of the subinterval l associated with the ith concretisation of the histogram-valued variable X and Y, respectively. $\bar{c}_{Xl}$ and $\bar{c}_{Yl}$, $\bar{r}_{Xl}$ and $\bar{r}_{Yl}$ are the centres and half ranges of the subinterval l associated with the Barycentric Histogram of X and Y, respectively.

Remark 2.18. The notation associated with the symbolic linear correlation chosen (r' and not r) aims at avoiding a possible confusion with the r used to refer to the half-range of a subinterval.

2.8.4.2 Mathematical Properties

This symbolic correlation coefficient is equipped with some of the usual mathematical properties associated with correlation coefficients for classical data. These properties are stated and proven below:

1. this measure is symmetric: $r'(X,Y) = r'(Y,X)$

   Proof. We start by noticing that:

$$s_X s_Y =$$

$$\begin{aligned}
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{Xl}\left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2\right]} \times \\
&\times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{Yl}\left[(c_{Y(i)l} - \bar{c}_{Yl})^2 + \frac{1}{3}(r_{Y(i)l} - \bar{r}_{Yl})^2\right]} = \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{Yl}\left[(c_{Y(i)l} - \bar{c}_{Yl})^2 + \frac{1}{3}(r_{Y(i)l} - \bar{r}_{Yl})^2\right]} \times \\
&\times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{Xl}\left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2\right]}
\end{aligned}$$

$$= s_Y s_X$$

   Moreover, since:

$$(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) = (c_{Y(i)l} - \bar{c}_{Yl})(c_{X(i)l} - \bar{c}_{Xl})$$

$$(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl}) = (r_{Y(i)l} - \bar{r}_{Yl})(r_{X(i)l} - \bar{r}_{Xl})$$

   We have the following result:

$$\begin{aligned}
cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l\left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l\left[(c_{Y(i)l} - \bar{c}_{Yl})(c_{X(i)l} - \bar{c}_{Xl}) + \frac{1}{3}(r_{Y(i)l} - \bar{r}_{Yl})(r_{X(i)l} - \bar{r}_{Xl})\right] \\
&= cov(Y,X)
\end{aligned} \tag{2.29}$$

Finally, we have that:

$$r'(X,Y) = \frac{cov(X,Y)}{s_X s_Y} = \frac{cov(Y,X)}{s_Y s_X} = r'(Y,X)$$

□

2. this measure is invariant under changes in location in the two histogram-valued variables: $r'(X+\alpha, Y+\beta) = r'(X,Y)$

   Proof. Recall that the operation of addition between a histogram-valued variable and a real value will only lead to a translation of the empirical distribution, that is, the half ranges remain unchanged whereas the centres are given by $c_{X(i)l} + \alpha$ and $c_{Y(i)l} + \beta$. Therefore, the centres of the barycentric histograms are given by $\frac{1}{n}\sum_{i=1}^{n}(c_{X(i)l} + \alpha) = \bar{c}_{Xl} + \alpha$ and $\frac{1}{n}\sum_{i=1}^{n}(c_{Y(i)l} + \beta) = \bar{c}_{Yl} + \beta$. We have that for each subinterval l:

$$(c_{X(i)l} + \alpha - (\bar{c}_{Xl} + \alpha))(c_{Y(i)l} + \beta - (\bar{c}_{Yl} + \beta)) = (c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl})$$

$$(c_{X(i)l} + \alpha - (\bar{c}_{Xl} + \alpha))^2 = (c_{X(i)l} - \bar{c}_{Xl})^2$$

$$(c_{Y(i)l} + \beta - (\bar{c}_{Yl} + \beta))^2 = (c_{Y(i)l} - \bar{c}_{Yl})^2$$

   Therefore, considering the expression 2.28:

$$r'(X+\alpha, Y+\beta) = \frac{cov(X+\alpha, Y+\beta)}{s_{X+\alpha} s_{Y+\beta}} = \frac{cov(X,Y)}{s_X s_Y} = r'(X,Y)$$

□

3. this measure is invariant under changes in scale in the two histogram-valued variables: assuming $\alpha, \beta \geq 0$, $r'(\alpha X, \beta Y) = r'(X,Y)$

   Remark 2.19. Note that, for $\alpha < 0$ or $\beta < 0$, we would obtain negative half-ranges. Therefore, it would not be a quantile function.

   Proof. Once again, we start by analysing the changes produced with regard to the centres and half-ranges. The centres are given by $\alpha c_{X(i)l}$ and $\beta c_{Y(i)l}$. Therefore, the centres of the barycentric histograms are given by $\frac{1}{n}\sum_{i=1}^{n}(\alpha c_{X(i)l}) = \alpha \bar{c}_{Xl}$ and $\frac{1}{n}\sum_{i=1}^{n}(\beta c_{Y(i)l}) = \beta \bar{c}_{Yl}$. We have that:

$$(\alpha c_{X(i)l} - \alpha\bar{c}_{Xl})(\beta c_{Y(i)l} - \beta\bar{c}_{Yl}) = \alpha(c_{X(i)l} - \bar{c}_{Xl})\beta(c_{Y(i)l} - \bar{c}_{Yl})$$
$$(\alpha c_{X(i)l} - \alpha\bar{c}_{Xl})^2 = \alpha^2(c_{X(i)l} - \bar{c}_{Xl})^2$$
$$(\beta\, c_{Y(i)l} - \beta\bar{c}_{Yl})^2 = \beta^2(c_{Y(i)l} - \bar{c}_{Yl})^2$$

Similarly, this reasoning can be applied to the half ranges, since they are affected in the same way as the centres. Therefore, we have that the expression of $cov(\alpha X, \beta Y)$ can be simplified:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[\alpha(c_{X(i)l} - \bar{c}_{Xl})\beta(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}\alpha(r_{X(i)l} - \bar{r}_{Xl})\beta(r_{Y(i)l} - \bar{r}_{Yl})\right] \\
&= \alpha\beta\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})(c_{Y(i)l} - \bar{c}_{Yl}) + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})(r_{Y(i)l} - \bar{r}_{Yl})\right] \\
&= \alpha\beta cov(X, Y)
\end{aligned} \tag{2.30}$$

Moreover,

$$\begin{aligned}
s_{\alpha X} &= \sqrt{var(\alpha X)} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[\alpha^2(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}\alpha^2(r_{X(i)l} - \bar{r}_{Xl})^2\right]} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l\alpha^2 \left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2\right]} \\
&= \alpha\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_l \left[(c_{X(i)l} - \bar{c}_{Xl})^2 + \frac{1}{3}(r_{X(i)l} - \bar{r}_{Xl})^2\right]} \\
&= \alpha s_X
\end{aligned} \tag{2.31}$$

Using a similar reasoning we can prove that:

$$s_{\beta Y} = \beta s_Y \tag{2.32}$$

Finally, we obtain:

$$r'(\alpha X, \beta Y) = \frac{cov(\alpha X, \beta Y)}{s_{\alpha X}s_{\beta Y}} = \frac{\alpha\beta cov(X, Y)}{\alpha s_X \beta s_Y} = \frac{cov(X, Y)}{s_X s_Y} = r'(X, Y)$$

□

Proposition 2.20. *The symbolic linear correlation coefficient between a histogram-valued variable X and a variable that is obtained by storing the symmetric histograms for each unit, $-X$, with regard to the y-axis, is -1 if the following conditions are verified:*

$$\Leftrightarrow \begin{cases} (\overline{c}_{X(m-l+1)} - \overline{c}_{Xl}) = (c_{X(i)(m-l+1)} - c_{X(i)l}) \\ (\overline{r}_{X(m-l+1)} + \overline{r}_{Xl}) = (r_{X(i)l} + r_{X(i)(m-l+1)}) \end{cases} \tag{2.33}$$

Proof. According to the expression 2.6, we have that:

$$c_{-X(i)l} = -c_{X(i)(m-l+1)}$$

$$\overline{c}_{-Xl} = \frac{1}{n}\sum_{i=1}^{n} c_{-X(i)l} = \frac{1}{n}\sum_{i=1}^{n} -c_{X(i)(m-l+1)} = -\overline{c}_{X(m-l+1)}$$

$$r_{-X(i)l} = r_{X(i)(m-l+1)}$$

$$\overline{r}_{-Xl} = \frac{1}{n}\sum_{i=1}^{n} r_{-X(i)l} = \frac{1}{n}\sum_{i=1}^{n} r_{X(i)(m-l+1)} = \overline{r}_{X(m-l+1)}$$

Therefore:

$$\begin{aligned} s_{-X} &= \sqrt{var(-X)} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{-X(i)l}\left[(c_{-X(i)l} - \overline{c}_{-Xl})^2 + \frac{1}{3}(r_{-X(i)l} - \overline{r}_{-Xl})^2\right]} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{-X(i)l}\left[(-c_{X(i)(m-l+1)} + \overline{c}_{X(m-l+1)})^2 + \frac{1}{3}(r_{X(i)(m-l+1)} - \overline{r}_{X(m-l+1)})^2\right]} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{-X(i)l}\left[(c_{X(i)(m-l+1)} - \overline{c}_{X(m-l+1)})^2 + \frac{1}{3}(r_{X(i)(m-l+1)} - \overline{r}_{X(m-l+1)})^2\right]} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{X(i)(m-l+1)}\left[(c_{X(i)(m-l+1)} - \overline{c}_{X(m-l+1)})^2 + \frac{1}{3}(r_{X(i)(m-l+1)} - \overline{r}_{X(m-l+1)})^2\right]} \end{aligned} \tag{2.34}$$

Let $l' = m - l + 1$, then:

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l'=1}^{m} p_{X(i)l'}\left[(c_{X(i)l'} - \overline{c}_{Xl'})^2 + \frac{1}{3}(r_{X(i)l'} - \overline{r}_{Xl'})^2\right]} = s_X$$

This means that the denominator of the symbolic correlation coefficient is given by:

$$s_{-X}s_X = s_Xs_X = s_X^2 = var(X) = cov(X, X)$$

Recall the formula of the symbolic covariance between X and $-X$.

$$cov(X,-X)=\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m}p_{X(i)l}\left[(c_{X(i)l}-\overline{c}_{Xl})(c_{-X(i)l}-\overline{c}_{-Xl})+\frac{1}{3}(r_{X(i)l}-\overline{r}_{Xl})(r_{-X(i)l}-\overline{r}_{-Xl})\right] \tag{2.35}$$

By exploring the symbolic correlation formula, it is possible to understand that expression 2.35 has to be equal to $-cov(X,X)$. As we will see, additional conditions must be imposed so that $r'(X,-X)=-1$. Consider the expression 2.35. The centres part of the equation may be rewritten as:

$$\begin{aligned}&(c_{X(i)l}-\overline{c}_{Xl})(c_{-X(i)l}-\overline{c}_{-Xl})=\\&(c_{X(i)l}-\overline{c}_{Xl})(-c_{X(i)(m-l+1)}+\overline{c}_{X(m-l+1)})\end{aligned} \tag{2.36}$$

The half ranges part of the equation may be rewritten as:

$$\begin{aligned}&(r_{X(i)l}-\overline{r}_{Xl})(r_{-X(i)l}-\overline{r}_{-Xl})=\\&(r_{X(i)l}-\overline{r}_{Xl})(r_{X(i)(m-l+1)}-\overline{r}_{X(m-l+1)})\end{aligned} \tag{2.37}$$

In order to obtain $cov(X,-X)=-cov(X,X)$, we must impose the following conditions. For $i\in\{1,...,n\}$ and $l\in\{1,...,m\}$,

$$\begin{cases}(c_{X(i)l}-\overline{c}_{Xl})=-(-c_{X(i)(m-l+1)}+\overline{c}_{X(m-l+1)})\\(r_{X(i)l}-\overline{r}_{Xl})=-(r_{X(i)(m-l+1)}-\overline{r}_{X(m-l+1)})\end{cases} \tag{2.38}$$

$$\Leftrightarrow\begin{cases}(\overline{c}_{X(m-l+1)}-\overline{c}_{Xl})=(c_{X(i)(m-l+1)}-c_{X(i)l})\\(\overline{r}_{X(m-l+1)}+\overline{r}_{Xl})=(r_{X(i)l}+r_{X(i)(m-l+1)})\end{cases} \tag{2.39}$$

Note that this is always true if we only have one unit. In other words, when only dealing with a histogram and the respective symmetric one.

The equations 2.36 and 2.37 can then be rewritten as:

$$(-1)(c_{X(i)l}-\overline{c}_{Xl})(c_{X(i)l}-\overline{c}_{Xl})=(-1)(c_{X(i)l}-\overline{c}_{Xl})^2 \tag{2.40}$$

$$(r_{X(i)l}-\overline{r}_{Xl})(-r_{X(i)l}+\overline{r}_{Xl})=(-1)(r_{X(i)l}-\overline{r}_{Xl})^2 \tag{2.41}$$

Moreover, expression 2.35 may be rewritten as:

$$
\begin{aligned}
cov(X,-X) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{X(i)l}\left[(-1)(c_{X(i)l}-\bar{c}_{Xl})^2+\frac{1}{3}(-1)(r_{X(i)l}-\bar{r}_{Xl})^2\right] \\
&= (-1)\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m} p_{X(i)l}\left[(c_{X(i)l}-\bar{c}_{Xl})^2+\frac{1}{3}(r_{X(i)l}-\bar{r}_{Xl})^2\right] \\
&= (-1)s_X^2
\end{aligned}
\tag{2.42}
$$

Finally, we can say that given a histogram-valued variable X and a variable that is obtained by storing the symmetric histograms for each unit, $-X$, such that $(\bar{c}_{X(m-l+1)} - \bar{c}_{Xl}) = (c_{X(i)(m-l+1)} - c_{X(i)l})$ and $(\bar{r}_{X(m-l+1)} + \bar{r}_{Xl}) = (r_{X(i)l} + r_{X(i)(m-l+1)})$:

$$
r'(X,-X) = \frac{cov(X,-X)}{s_X s_{-X}} = \frac{(-1)s_X^2}{s_X^2} = -1 \tag{2.43}
$$

□

Example 2.6. *To illustrate the previous proposition, Figures 2.8 and 2.9 display the scatter plot of two symbolic data sets. Although it seems that both show perfect inverse relation, the correlation values are not both equal to -1.*

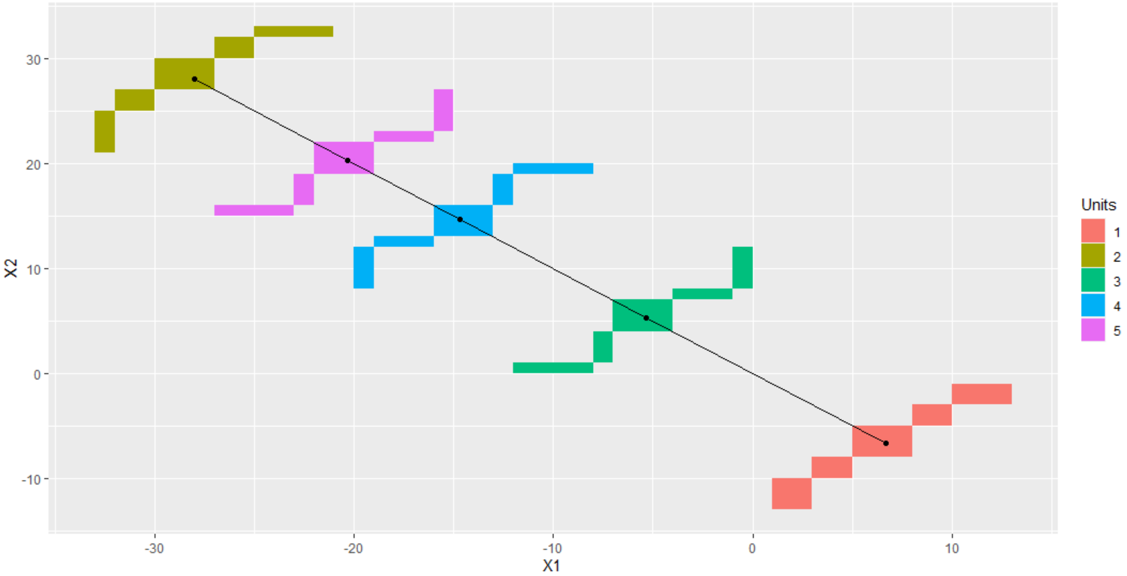


FIGURE 2.8: Scatter plot of a pair of histogram-valued variables with correlation value $r' = -1$.

Consider Tables 2.6 and 2.7 that represent the data used to obtain Figures 2.8 and 2.9, respectively.

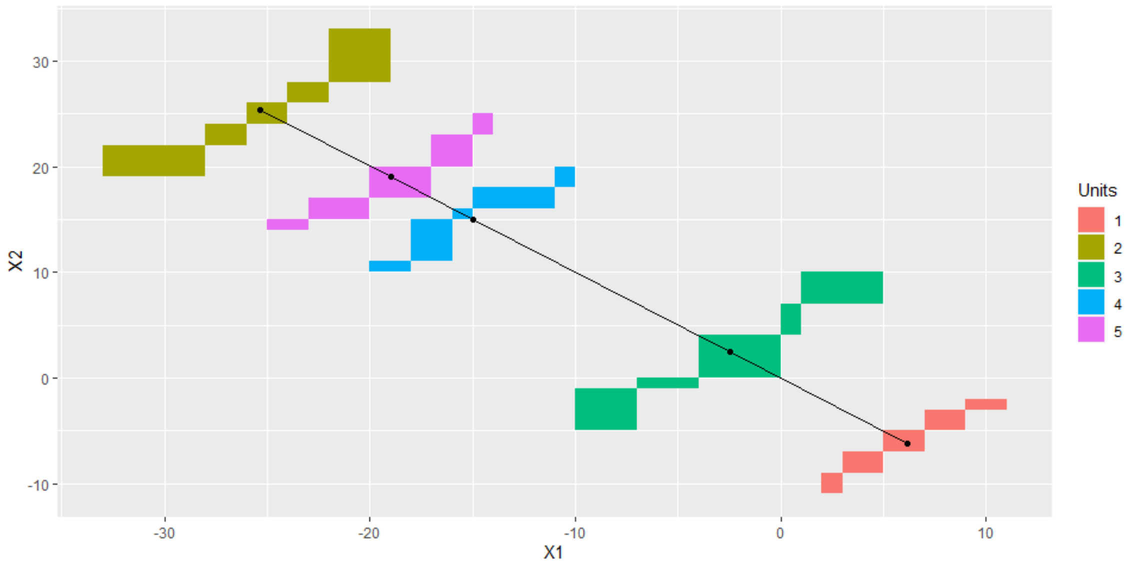


FIGURE 2.9: Scatter plot of a pair of histogram-valued variables with correlation value $r' = -0.9950$.

	X	$-X$
1	{ [1,3[,0.2; [3,5[,0.2; [5,8[,0.2; [8,10[,0.2; [10,13],0.2 }	{ [-13,-10[,0.2; [-10,-8[,0.2; [-8,-5[,0.2; [-5,-3[,0.2; [-3,-1],0.2 }
2	{ [-33,-32[,0.2; [-32,-30[,0.2; [-30,-27[,0.2; [-27,-25[,0.2; [-25,-21],0.2 }	{ [21,25[,0.2; [25,27[,0.2; [27,30[,0.2; [30,32[,0.2; [32,33],0.2 }
3	{ [-12,-8[,0.2; [-8,-7[,0.2; [-7,-4[,0.2;[-4,-1[,0.2; [-1,0],0.2 }	{ [0,1[,0.2; [1,4[,0.2; [4,7[,0.2;[7,8[,0.2; [8,12],0.2 }
4	{ [-20,-19[,0.2; [-19,-16[,0.2; [-16,-13[,0.2; [-13,-12[,0.2; [-12,-8],0.2 }	{ [8,12[,0.2; [12,13[,0.2; [13,16[,0.2; [16,19[,0.2; [19,20],0.2 }
5	{ [-27,-23[,0.2; [-23,-22[,0.2; [-22,-19[,0.2; [-19,-16[,0.2; [-16,-15],0.2 }	{ [15,16[,0.2; [16,19[,0.2; [19,22[,0.2; [22,23[,0.2; [23,27],0.2 }

TABLE 2.6: Symbolic data table of a pair of histogram-valued variables with correlation value $r' = -1$.

	Y	$-Y$
1	{ [2,3[,0.2; [3,5[,0.2; [5,7[,0.2; [7,9[,0.2; [9,11],0.2 }	{ [-11,-9[,0.2; [-9,-7[,0.2; [-7,-5[,0.2; [-5,-3[,0.2; [-3,-2],0.2 }
2	{ [-33,-28[,0.2; [-28,-26[,0.2; [-26,-24[,0.2; [-24,-22[,0.2; [-22,-19],0.2 }	{ [19,22[,0.2; [22,24[,0.2; [24,26[,0.2; [26,28[,0.2; [28,33],0.2 }
3	{ [-10,-7[,0.2; [-7,-4[,0.2; [-4,0[,0.2;[0,1[,0.2; [1,5],0.2 }	{ [-5,-1[,0.2; [-1,0[,0.2; [0,4[,0.2;[4,7[,0.2; [7,10],0.2 }
4	{ [-20,-18[,0.2; [-18,-16[,0.2; [-16,-15[,0.2; [-15,-11[,0.2; [-11,-10],0.2 }	{ [10,11[,0.2; [11,15[,0.2; [15,16[,0.2; [16,18[,0.2; [18,20],0.2 }
5	{ [-25,-23[,0.2; [-23,-20[,0.2; [-20,-17[,0.2; [-17,-15[,0.2; [-15,-14],0.2 }	{ [14,15[,0.2; [15,17[,0.2; [17,-20[,0.2; [20,23[,0.2; [23,25],0.2 }

TABLE 2.7: Symbolic data table of a pair of histogram-valued variables with correlation value $r' = -0.9950$.

Unit	Subinterval	$c_{X(i)l}$	$\bar{c}_{Xl}$	$r_{X(i)l}$	$\bar{r}_{Xl}$
1	1	2	-17	1	1.2
	2	4	-14.9	1	0.9
	3	6.5	-12.5	1.5	1.5
	4	9	-9.9	1	1.1
	5	11.5	-7.5	1.5	1.3
2	1	-32.5	-17	0.5	1.2
	2	-31	-14.9	1	0.9
	3	-28.5	-12.5	1.5	1.5
	4	-26	-9.9	1	1.1
	5	-23	-7.5	2	1.3
3	1	-10	-17	2	1.2
	2	-7.5	-14.9	0.5	0.9
	3	-5.5	-12.5	1.5	1.5
	4	-2.5	-9.9	1.5	1.1
	5	-0.5	-7.5	0.5	1.3
4	1	-19.5	-17	0.5	1.2
	2	-17.5	-14.9	1.5	0.9
	3	-14.5	-12.5	1.5	1.5
	4	-12.5	-9.9	0.5	1.1
	5	-10	-7.5	2	1.3
5	1	-25	-17	2	1.2
	2	-22.5	-14.9	0.5	0.9
	3	-20.5	-12.5	1.5	1.5
	4	-17.5	-9.9	1.5	1.1
	5	-15.5	-7.5	0.5	1.3

TABLE 2.8: Calculation of centres and half ranges of the histogram-valued variable X and of the Barycentric Histogram of X, regarding data with correlation value $r' = -1$.

According to the data, it is possible to calculate the values displayed in Tables 2.8 and 2.9. By observing these tables, it is evident that the data in Table 2.6, i.e., the data of a pair of histogram-valued variables with correlation value $r' = -1$ verify conditions 2.33.

It is also possible to calculate the values displayed in Tables 2.10 and 2.11. By observing these tables, it is evident that the data in Table 2.7, i.e., the data of a pair of histogram-valued variables with correlation value $r' = -0.9950$ do not verify conditions 2.33.

Example 2.7. *Consider once again the data associated with the Tables 2.3, 2.4 and 2.5. Figures 2.10, 2.11 and 2.12 not only show the scatter plots of the variables associated with the data, but also the correlation values.*

As expected, the correlation values are in agreement with the relations deduced previously. The data with strong direct relation, in Figure 2.10, shows a high positive correlation value. The data with strong inverse relation, in Figure 2.11, shows a high negative value of correlation. Finally,

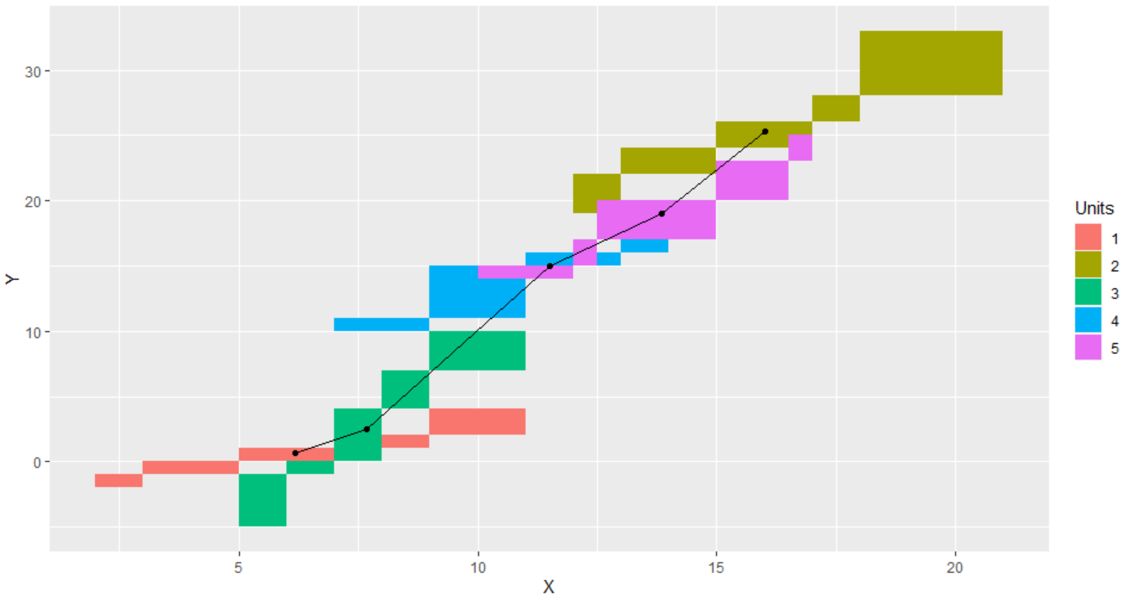


FIGURE 2.10: Scatter plot of a pair of histogram-valued variables with correlation value $r' = 0.9789$.

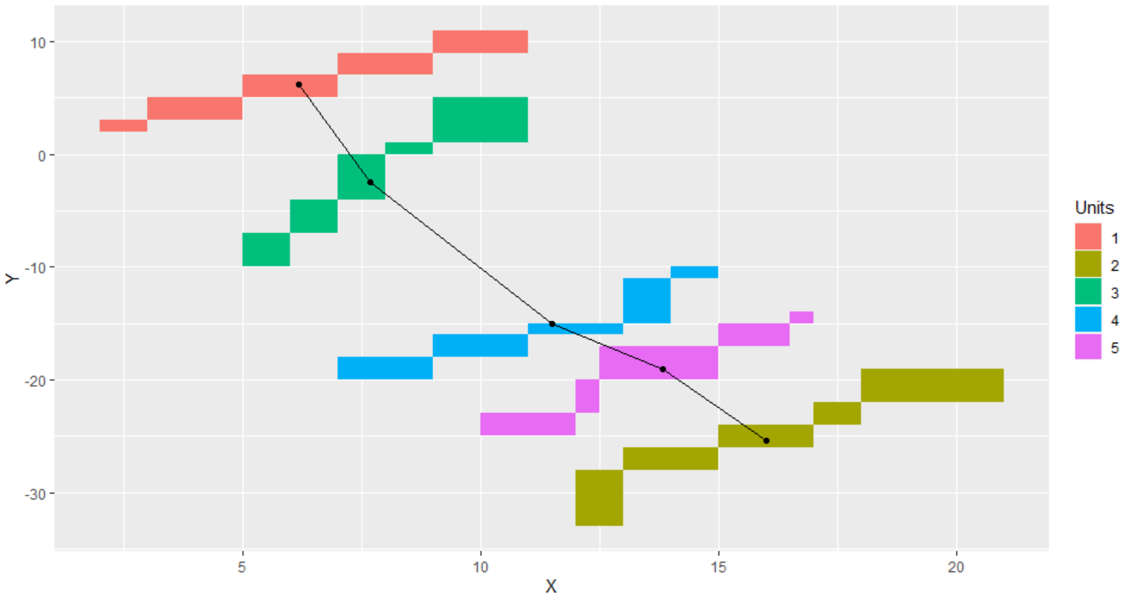


FIGURE 2.11: Scatter plot of a pair of histogram-valued variables with correlation value $r' = -0.9947$.

Unit	Subinterval	$c_{X(i)(m-l+1)}-$	$\bar{c}_{X(m-l+1)}-$	$r_{X(i)l}+$	$\bar{r}_{X(m-l+1)}+$
		$-c_{X(i)l}$	$-\bar{c}_{Xl}$	$+r_{X(i)(m-l+1)}$	$+\bar{r}_{Xl}$
1	1	-9.5	-9.5	2.5	2.5
	2	-5	-5	2	2
	3	0	0	3	3
	4	5	5	2	2
	5	9.5	9.5	2.5	2.5
2	1	-9.5	-9.5	2.5	2.5
	2	-5	-5	2	2
	3	0	0	3	3
	4	5	5	2	2
	5	9.5	9.55	2.5	2.5
3	1	-9.5	-9.5	2.5	2.5
	2	-5	-5	2	2
	3	0	0	3	3
	4	5	5	2	2
	5	9.5	9.55	2.5	2.5
4	1	-9.5	-9.5	2.5	2.5
	2	-5	-5	2	2
	3	0	0	3	3
	4	5	5	2	2
	5	9.5	9.55	2.5	2.5
5	1	-9.5	-9.5	2.5	2.5
	2	-5	-5	2	2
	3	0	0	3	3
	4	5	5	2	2
	5	9.5	9.55	2.5	2.5

Table 2.9: Verification of conditions 2.33 on a pair of histogram-valued variables with correlation value $r' = -1$.

the data that does not display clear direct or inverse relation, in Figure 2.12, discloses the least absolute value of correlation. In fact, this value is very close to 0.

Unit	Subinterval	$c_{Y(i)l}$	$\bar{c}_{Yl}$	$r_{Y(i)l}$	$\bar{r}_{Yl}$
1	1	2.5	-15.9	0.5	1.3
	2	4	-13.4	1	1.2
	3	6	-11	1	1.2
	4	8	-8.7	1	1.1
	5	10	-6.5	1	1.1
2	1	-30.5	-15.9	2.5	1.3
	2	-27	-13.4	1	1.2
	3	-25	-11	1	1.2
	4	-23	-8.7	1	1.1
	5	-20.5	-6.5	1.5	1.1
3	1	-8.5	-15.9	1.5	1.3
	2	-5.5	-13.4	1.5	1.2
	3	-2	-11	2	1.2
	4	0.5	-8.7	0.5	1.1
	5	3	-6.5	2	1.1
4	1	-19	-15.9	1	1.3
	2	-17	-13.4	1	1.2
	3	-15.5	-11	0.5	1.2
	4	-13	-8.7	2	1.1
	5	-10.5	-6.5	0.5	1.1
5	1	-24	-15.9	1	1.3
	2	-21.5	-13.4	1.5	1.2
	3	-18.5	-11	1.5	1.2
	4	-16	-8.7	1	1.1
	5	-14.5	-6.5	0.5	1.1

TABLE 2.10: Calculation of centres and half ranges of the histogram-valued variable Y and of the Barycentric Histogram of Y, regarding data with correlation value $r' = -0.9947$.

Unit	Subinterval	$c_{Y(i)(m-l+1)}-$ $-c_{Y(i)l}$	$\bar{c}_{Y(m-l+1)}-$ $-\bar{c}_{Yl}$	$r_{Y(i)l}+$ $+r_{Y(i)(m-l+1)}$	$\bar{r}_{Y(m-l+1)}+$ $+\bar{r}_{Yl}$
1	1	-7.5	-9.4	1.5	2.4
	2	-4	-4.7	2	2.3
	3	0	0	2	2.4
	4	4	4.7	2	2.3
	5	7.5	9.4	1.5	2.4
2	1	-10.5	-9.4	4	2.4
	2	-4	-4.7	2	2.3
	3	0	0	2	2.4
	4	4	4.7	2	2.3
	5	10.5	9.4	4	2.4
3	1	-11.5	-9.4	3.5	2.4
	2	-6	-4.7	2	2.3
	3	0	0	4	2.4
	4	6	4.7	2	2.3
	5	11.5	9.4	3.5	2.4
4	1	-8.5	-9.4	1.5	2.4
	2	-4	-4.7	3	2.3
	3	0	0	1	2.4
	4	4	4.7	3	2.3
	5	8.5	9.4	1.5	2.4
5	1	-9.5	-9.4	1.5	2.4
	2	-5.5	-4.7	2.5	2.3
	3	0	0	3	2.4
	4	5.5	4.7	2.5	2.3
	5	9.5	9.4	1.5	2.4

TABLE 2.11: Verification of conditions 2.33 on a pair of histogram-valued variables with correlation value $r' = -0.9950$.

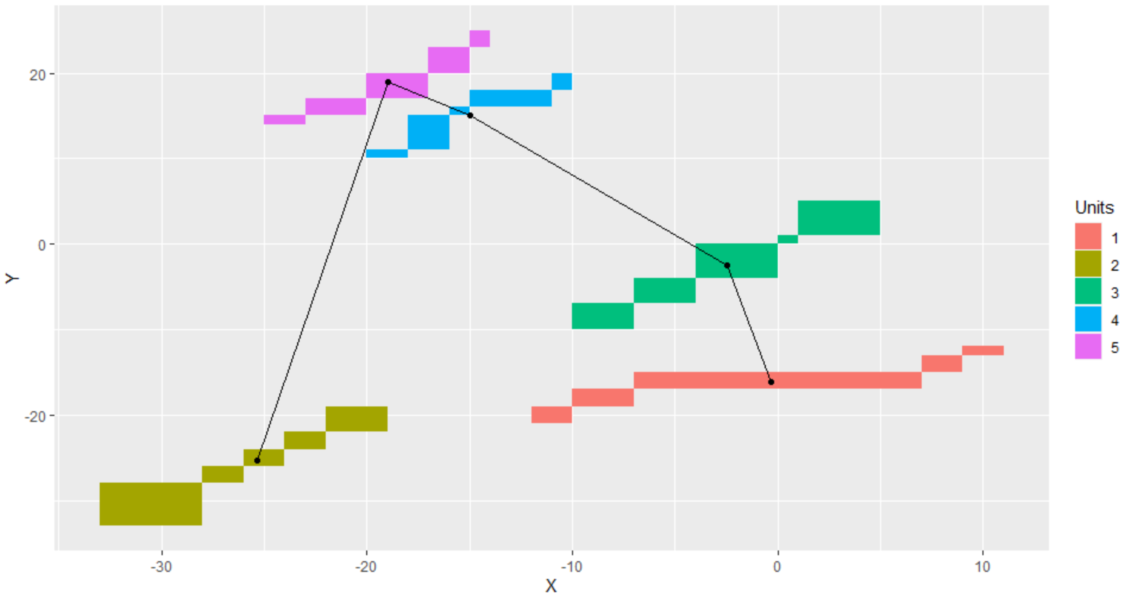


FIGURE 2.12: Scatter plot of a pair of histogram-valued variables with correlation value $r' = -0.0567$.

2.9 Linear Combination of Histogram-Valued Variables

We aim at performing linear discriminant analysis. In other words, we aim at finding a linear combination of variables that distinguishes two or more classes. For that reason, it is important to adapt the classical definition of linear combination so that it becomes appropriate for symbolic variables.

2.9.1 Assumptions

The definition of a linear discriminant function revolves around the maximisation of the Between Inertia/Within Inertia ratio. This maximisation allows obtaining the optimal parameters for the linear discriminant function. The use of the Mallows distance grants this optimisation step.

However, the definitions given in this chapter take into account some requirements. Therefore, to develop this method we must impose these requirements as a way of using the concepts and *formulae* disclosed. Consider the following assumptions:

1. All quantile functions involved need to have an equal number of pieces. In other words, all histograms involved need to have an equal number of subintervals;

2. The domain of each piece has to be the same for all involved functions (quantile functions and those that represent the symmetric histograms);

3. For all histograms involved, $h = \{I_{h1}, p_{h1}; I_{h2}, p_{h2}; ...; I_{hm}, p_{hm}\}$, the weights p_{hl} have to verify the condition:

$$\underset{l \in \{1,...,m\}}{\forall} p_{hl} = p_{h(m-l+1)}$$

Note that, when the rewriting process, displayed in Section 2.4, is performed, the obtained data already verifies these conditions.

2.9.2 Linear Combination

The adjustment to the classical definition of linear combination is mainly due to the fact that when multiplying a quantile function by a negative number, a quantile function is not obtained, since it is not obtained a non-decreasing function. As a first attempt to fix this problem, in [4], it was considered a linear combination of quantile functions with positive parameters:

$$\Psi_{i(p+1)}(q) = a_1\Psi_{i1}(q) + ... + a_p\Psi_{ip}(q),$$

where $a_1, ..., a_p \in \mathbb{R}^+$. However, in this scenario, we are forcing a direct linear relation between $\Psi_{i(p+1)}(q)$ and $\Psi_{ij}(q)$, with $j \in \{1, ..., p\}$.

Since the space of quantile functions is a semi-vector space, the following definition solves a critical problem. For this, it uses both the quantile function of the observed histograms, together with those of the corresponding symmetric histograms.

Definition 2.21. Consider the histogram valued variables $X_1, X_2, \ldots, X_p$ with quantile functions for each unit i $\Psi_{i1}(q), \Psi_{i2}(q), ..., \Psi_{ip}(q)$, with $q \in [0, 1]$, and the quantile functions that represent the respective symmetric histograms $-\Psi_{i1}(1-q), -\Psi_{i2}(1-q), ..., -\Psi_{ip}(1-q)$. The linear combination of $X_1, X_2, ..., X_p$, presented in [4], is a new histogram-valued variable, designated by $X_{(p+1)}$, where each quantile function of unit i can be written as:

$$\Psi_{i(p+1)}(q) = \sum_{j=1}^{p} a_j\Psi_{ij}(q) - \sum_{j=1}^{p} b_j\Psi_{ij}(1-q), \tag{2.44}$$

where $q \in [0, 1]$, $a_j, b_j \geq 0$ and $j \in \{1, ..., p\}$.

This definition of a linear combination of histogram-valued variables does not force a direct linear relationship in view of the fact that it uses both the quantile functions of the histogram-valued variables and the quantile functions of the corresponding symmetric histograms, as it was stated in [4]. Therefore, this definition is an improvement from the previous one.

Recall, once again, that interval-valued variables may be considered a particular case of the histogram-valued variables. When analysing the interval-valued variables, the previous expressions become easier to look over. Following the reasoning developed in [6], the quantile function associated with the ith concretisation of the interval-valued variable X_j is:

$$\Psi_{ij}(q) = c_{ij} + r_{ij}(2q-1), q \in [0, 1] \tag{2.45}$$

The quantile function that represents the symmetric interval is then:

$$\begin{aligned} -\Psi_{ij}(1-q) &= -c_{ij} - r_{ij}(2(1-q)-1) \\ &= -c_{ij} - r_{ij}(-2q+1) \\ &= -c_{ij} + r_{ij}(2q-1) \end{aligned} \tag{2.46}$$

The linear combination of $X_1, X_2, ..., X_p$ is a new interval-valued variable $X_{(p+1)}$ where the quantile function of unit i can be written as:

$$\begin{aligned} \Psi_{i(p+1)}(q) &= \sum_{j=1}^{p} a_j(c_{ij} + r_{ij}(2q-1)) + \sum_{j=1}^{p} b_j(-c_{ij} + r_{ij}(2q-1)) \\ &= \sum_{j=1}^{p}(a_j - b_j)c_{ij} + \sum_{j=1}^{p}(a_j + b_j)r_{ij}(2q-1) \end{aligned} \tag{2.47}$$

Furthermore, if we consider the case where the range of the interval is null, that is, $X_j(i) = h_{ij} = I_{ij} = [x_{ij}]$. Then $r_{ij} = 0$ and $c_{ij} = x_{ij}$ and the equation becomes:

$$\begin{aligned} \Psi_{i(p+1)}(q) &= \sum_{j=1}^{p}(a_j - b_j)c_{ij} \\ &= \sum_{j=1}^{p}(a_j - b_j)x_{ij} \\ &= \sum_{j=1}^{p} d_j x_{ij} \end{aligned} \tag{2.48}$$

which is consistent with the classical definition of linear combination.

Chapter 3

Linear Discriminant Analysis for Histogram-Valued Variables

At this moment we are equipped with the necessary tools to develop the linear discriminant model for histogram-valued variables.

In this section, it is introduced crucial concepts such as the score quantile function and the barycentric scores. Moreover, we present an explanation of the optimisation and classification steps performed for the two *a priori* classes case scenario. Finally, we reveal the three approaches explored to address the multi-class classification problem as well as the optimisation and classification steps required.

3.1 Linear Discriminant Function

The linear discriminant function proposed in [1] defines a score, $S(i)$, for each unit i, and it is given by the linear combination of p explanatory histogram-valued variables. The score is then used to classify that unit. This classification is based on the distance between the unit's score and the barycentric score of each *a priori* class.

We make use of the definition of linear combination of histogram-valued variables proposed in [4] and the assumptions disclosed in Section 2.9.1.

Definition 3.1 (Score Quantile Function)**.** Consider the histogram-valued variables X_1, X_2 $..., X_p$ with quantile functions for each unit i $\Psi_{i1}(q), \Psi_{i2}(q), ..., \Psi_{ip}(q)$, with $q \in [0,1]$, and the quantile functions that represent the respective symmetric histograms $-\Psi_{i1}(1-q), -\Psi_{i2}(1-q), ..., -\Psi_{ip}(1-q)$. The score of unit i introduced in [1] is the quantile function:

$$\Psi_{S(i)}(q) = \sum_{j=1}^{p} a_j \Psi_{ij}(q) - \sum_{j=1}^{p} b_j \Psi_{ij}(1-q), \tag{3.1}$$

where $q \in [0,1]$, $a_j, b_j \geq 0$ and $j \in \{1, ..., p\}$.

Let S denote the score's histogram-valued variable.

For each subinterval l and $q \in [w_{(l-1)}, w_l[$, the score quantile function of unit i is given by:

$$\sum_{j=1}^{p} \left(a_j c_{ijl} - b_j c_{ij(m-l+1)} \right) + \left(\frac{2(q - w_{(l-1)})}{w_l - w_{(l-1)}} - 1 \right) \sum_{j=1}^{p} \left(a_j r_{ijl} + b_j r_{ij(m-l+1)} \right) \tag{3.2}$$

Definition 3.2 ((Global) Barycentric Score)**.** The global barycentric score (also simply mentioned as Barycentric Score), $\overline{\Psi_S}(q)$, is the mean of the quantile functions that represent individual scores. For subinterval l, given $q \in [w_{(l-1)}, w_l[$,

$$\overline{\Psi_S}(q) = \sum_{j=1}^{p} \left(a_j \bar{c}_{jl} - b_j \bar{c}_{j(m-l+1)} \right) + \left(\frac{2(q - w_{(l-1)})}{w_l - w_{(l-1)}} - 1 \right) \sum_{j=1}^{p} \left(a_j \bar{r}_{jl} + b_j \bar{r}_{j(m-l+1)} \right), \tag{3.3}$$

where $\bar{c}_{jl}$ and $\bar{r}_{jl}$ are the means of the centres and the means of the half ranges, considering all units, of the subinterval l for variable j, respectively [1].

Definition 3.3 (Barycentric Score of a Group)**.** The Barycentric Score of a Group u of units, $\overline{\Psi_{S_u}}(q)$, is the mean of the quantile functions that represent individual scores within group u. For each sub-interval l, given $q \in [w_{(l-1)}, w_l[$,

$$\overline{\Psi_{S_u}}(q) = \sum_{j=1}^{p} \left(a_j \bar{c}_{jlu} - b_j \bar{c}_{j(m-l+1)u} \right) + \left(\frac{2(q - w_{(l-1)})}{w_l - w_{(l-1)}} - 1 \right) \sum_{j=1}^{p} \left(a_j \bar{r}_{jlu} + b_j \bar{r}_{j(m-l+1)u} \right), \tag{3.4}$$

where $\bar{c}_{jlu}$ and $\bar{r}_{jlu}$ are the means of the centres and the means of the half ranges of the units in group u and sub-interval l for variable j, respectively [1].

Theorem 3.4. *Let $\Psi_{S(i)}$ be the score quantile function of unit i, considering the histogram-valued variables $X_1, X_2, \ldots, X_p$ with quantile functions for each unit i $\Psi_{i1}, \Psi_{i2}, \ldots, \Psi_{ip}$. Moreover, let $\overline{\Psi_S}$ be the barycentric score. In [1], it is shown that the sum of the squared Mallows distance between $\Psi_{S(i)}$ and $\overline{\Psi_S}$ can be rewritten as:*

$$\begin{aligned}\sum_{i=1}^{n} D_M^2(\Psi_{S(i)},\overline{\Psi_S}) &= \gamma^T T\gamma \\ &= \sum_{u=1}^{s} n_u D_M^2(\overline{\Psi_{S_u}},\overline{\Psi_S}) + \sum_{u=1}^{s}\sum_{S(i)\in G_u} D_M^2(\Psi_{S(i)},\overline{\Psi_{S_u}}) \\ &= \gamma^T B\gamma + \gamma^T W\gamma\end{aligned} \tag{3.5}$$

where $\gamma = (a_1, b_1, ..., a_p, b_p)$, i.e., it is a $2p \times 1$ (column) vector, n_u is the cardinality of the group u, $T = [T_{de}]$ is the (symmetric) matrix of the total Sums of Squares and Cross-Products (SSCP) for the p histogram-valued variables, $B = [B_{de}]$ and $W = [W_{de}]$ are the matrices of the SSCP between-groups and within-groups, respectively. The three mentioned matrices are symmetric of order $2p$. The elements are given by the following expressions:

$$T_{de} = \begin{cases} \sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l\left(\tilde{c}_{i\frac{d+1}{2}l}\tilde{c}_{i\frac{e+1}{2}l} + \frac{1}{3}\tilde{r}_{i\frac{d+1}{2}l}\tilde{r}_{i\frac{e+1}{2}l}\right) & \textit{if d,e are odd} \\ \sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l\left(\tilde{c}_{i\frac{d}{2}(m-l+1)}\tilde{c}_{i\frac{e}{2}(m-l+1)} + \frac{1}{3}\tilde{r}_{i\frac{d}{2}(m-l+1)}\tilde{r}_{i\frac{e}{2}(m-l+1)}\right) & \textit{if d,e are even} \\ \sum\limits_{i=1}^{n}\sum\limits_{l=1}^{m} p_l\left(-\tilde{c}_{i\frac{d}{2}l}\tilde{c}_{i\frac{e+1}{2}(m-l+1)} + \frac{1}{3}\tilde{r}_{i\frac{d}{2}l}\tilde{r}_{i\frac{e+1}{2}(m-l+1)}\right) & \textit{if d is even, e is odd} \end{cases} \tag{3.6}$$

where $\tilde{c}_{ijl} = c_{ijl} - \bar{c}_{jl}$ and $\tilde{r}_{ijl} = r_{ijl} - \bar{r}_{jl}$, $j \in \{\frac{d+1}{2}, \frac{e+1}{2}, \frac{d}{2}, \frac{e}{2}\}$.

$$B_{de} = \begin{cases} \sum\limits_{u=1}^{s} n_u \sum\limits_{l=1}^{m} p_l\left(\breve{c}_{\frac{d+1}{2}lu}\breve{c}_{\frac{e+1}{2}lu} + \frac{1}{3}\breve{r}_{\frac{d+1}{2}lu}\breve{r}_{\frac{e+1}{2}lu}\right) & \textit{if d,e are odd} \\ \sum\limits_{u=1}^{s} n_u \sum\limits_{l=1}^{m} p_l\left(\breve{c}_{\frac{d}{2}(m-l+1)u}\breve{c}_{\frac{e}{2}(m-l+1)u} + \frac{1}{3}\breve{r}_{\frac{d}{2}(m-l+1)u}\breve{r}_{\frac{e}{2}(m-l+1)u}\right) & \textit{if d,e are even} \\ \sum\limits_{u=1}^{s} n_u \sum\limits_{l=1}^{m} p_l\left(-\breve{c}_{\frac{d}{2}lu}\breve{c}_{\frac{e+1}{2}(m-l+1)u} + \frac{1}{3}\breve{r}_{\frac{d}{2}lu}\breve{r}_{\frac{e+1}{2}(m-l+1)u}\right) & \textit{if d is even, e is odd} \end{cases} \tag{3.7}$$

where $\breve{c}_{jlu} = \bar{c}_{jl} - \bar{c}_{jlu}$ and $\breve{r}_{jlu} = \bar{r}_{jl} - \bar{r}_{jlu}$, $j \in \{\frac{d+1}{2}, \frac{e+1}{2}, \frac{d}{2}, \frac{e}{2}\}$.

$$W_{de} = \begin{cases} \sum\limits_{u=1}^{s}\sum\limits_{i\in G'_u}\sum\limits_{l=1}^{m} p_l\left(\tilde{c}_{i\frac{d+1}{2}lu}\tilde{c}_{i\frac{e+1}{2}lu} + \frac{1}{3}\tilde{r}_{i\frac{d+1}{2}lu}\tilde{r}_{i\frac{e+1}{2}lu}\right) & \textit{if d,e are odd} \\ \sum\limits_{u=1}^{s}\sum\limits_{i\in G'_u}\sum\limits_{l=1}^{m} p_l\left(\tilde{c}_{i\frac{d}{2}(m-l+1)u}\tilde{c}_{i\frac{e}{2}(m-l+1)u} + \frac{1}{3}\tilde{r}_{i\frac{d}{2}(m-l+1)u}\tilde{r}_{i\frac{e}{2}(m-l+1)u}\right) & \textit{if d,e are even} \\ \sum\limits_{u=1}^{s}\sum\limits_{i\in G'_u}\sum\limits_{l=1}^{m} p_l\left(-\tilde{c}_{i\frac{d}{2}lu}\tilde{c}_{i\frac{e+1}{2}(m-l+1)u} + \frac{1}{3}\tilde{r}_{i\frac{d}{2}lu}\tilde{r}_{i\frac{e+1}{2}(m-l+1)u}\right) & \textit{if d is even, e is odd} \end{cases} \tag{3.8}$$

where $\tilde{c}_{ijlu} = c_{ijl} - \bar{c}_{jlu}$ and $\tilde{r}_{ijl} = r_{ijl} - \bar{r}_{jlu}$ and $G'_u = \{i : \forall_j\, X_j(i) \in G_u\}$, $j \in \{\frac{d+1}{2}, \frac{e+1}{2}, \frac{d}{2}, \frac{e}{2}\}$.

The detailed construction and proof of the *formulae* of the matrices mentioned above can be found in [1].

3.2 Classification in Two *A Priori* Groups

3.2.1 Optimisation

Consider p histogram-valued variables, $X_1, X_2, \ldots, X_p$. Let $\gamma = (a_1, b_1, ..., a_p, b_p)$, i.e, γ is the parameter vector that defines the linear discriminant function. The optimal parameter vector is estimated such that the ratio, λ, regarding the variability between groups and the variability within groups is maximum [1]. The measure λ represents how well the parameter vector is useful to separate the data in classes. In mathematical formulation, it can be written as the following optimisation problem:

$$\gamma^* = \arg\max_{\gamma} \lambda = \arg\max_{\gamma} \frac{\gamma^\mathsf{T} B \gamma}{\gamma^\mathsf{T} W \gamma} \tag{3.9}$$

subject to:

$$\gamma \geq 0$$

This is a constrained quadratic optimisation problem. It is subjected to non-negativity constraints on the parameters. As a consequence, it becomes a hard problem to solve. Note that this restriction does not force a direct linear relationship between the variables and scores, because the score quantile function definition uses both the quantile functions of the histograms and the quantile functions of the corresponding symmetric histograms.

In [18], it is investigated formulations for the given problem. The major disadvantage is connected to the algorithmic aspect. Based on the work developed, it is possible to derive semi-definite programming relaxations in the interest of finding good upper bounds to this problem. Semi-definite programming is a field concerned with the optimisation of a linear objective function over the intersection of the cone of positive semi-definite matrices with an affine space [19].

Remark 3.5. M is a positive semi-definite matrix $\iff$ $\mathbf{x}^\mathsf{T} M \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$

The relaxations are then used in the global optimisation approach in order to prove optimality.

As it is possible to understand, for $\lambda < 1$, we have that $\gamma^{\mathsf{T}} B\gamma < \gamma^{\mathsf{T}} W\gamma$. In this case, data does not display well-defined groups, because the distance within groups is higher than the one between them. Since data does not seem to have a clusterable structure, it stands to reason that the model developed will not perform well.

3.2.2 Classification

Once the optimal parameter vector is obtained, the linear discriminant function is defined. The general idea behind the classification is to assign a given unit i to the group that is at the minimum distance. For that purpose, it is calculated the score quantile function associated with each unit i and the barycentric scores of all the groups of units, provided that a group is defined by the *a priori* classes. Using the Mallows distance, it is then possible to assign the unit to the group such that its barycentric score is closest to the unit's score [1].

Definition 3.6. Consider the data divided in two *a priori* groups G_1 and G_2, with barycentric scores for each group $\overline{\Psi_{S_1}}$ and $\overline{\Psi_{S_2}}$, respectively. Let $\Psi_{S_{(i)}}$ be the score quantile function of unit i. The classification performed by the developed model for unit i is given by:

$$\begin{cases} G_1 & \text{if } D_M^2(\overline{\Psi_{S_1}}, \Psi_{S(i)}) < D_M^2(\overline{\Psi_{S_2}}, \Psi_{S(i)}) \\ G_2 & \text{if } D_M^2(\overline{\Psi_{S_1}}, \Psi_{S(i)}) > D_M^2(\overline{\Psi_{S_2}}, \Psi_{S(i)}) \\ sample(G_1, G_2) & \text{otherwise} \end{cases} \tag{3.10}$$

It could be the case that a given unit is equally spaced with respect to the barycentric scores of both groups G_1 and G_2. In such a case, $sample(G_1, G_2)$ is performed. It stands for a random sample of size one between both groups without replacement. This computation introduces randomness associated with the classification process.

This definition differs from the one presented in [1] regarding the $sample$ operation. In [1], unit i is assigned to group G_1 if $D_M^2(\overline{\Psi_{S_1}}, \Psi_{S(i)}) < D_M^2(\overline{\Psi_{S_2}}, \Psi_{S(i)})$, otherwise it is assigned to group G_2.

Since this is a classification problem, it could be the case that unbalanced groups would have a great impact on the models' performance. However, in simulations conducted in [1], it was concluded that the disturbance caused by unbalanced groups is neglectable.

3.3 Multi-class Classification

Heretofore the classification problem with two *a priori* groups has been dealt with. Although the multi-class classification problem is not immediately solved, the binary classification is a step towards the solution. When considering more than two *a priori* classes, there are two ideas that arise:

1. Divide the multi-class classification dataset into several binary classification subproblems. In this case, identifying the best multi-class classifier involves finding the best binary classifiers. In other words, we are using the already existing binary class classifiers. Concerning this approach, there are two well-known types of multi-class classification techniques:
    (a) One-Versus-One (OVO);
    (b) One-Versus-All (OVA);
2. Define several linear discriminant functions with the condition that they must be uncorrelated with each other. In this scenario, a good multi-class classifier is built on the search for as many useful discriminant functions as possible. This idea is referenced as Consecutive Linear Discriminant Functions (CLDF).

3.3.1 Multiple Binary Classification Subproblems

3.3.1.1 One-Versus-One (OVO)

Consider a classification problem with s groups, $\{G_1, G_2, ..., G_s\}$. One-Versus-One (OVO) is a technique that decomposes a problem with s groups into $\frac{s(s-1)}{2}$ binary subproblems, where each problem is addressed by a binary classifier [20]. Note that the number of binary subproblems corresponds to $\binom{s}{2}$, i.e., the possible combinations of 2 groups between the s groups. Each binary classification is performed only over the data related with the two groups involved. In other words, we are performing the binary classification over a subset of the original data set. The final assignment of which group a unit i belongs to is performed by using the majority vote.

3.3.1.2 One-Versus-All (OVA)

Suppose a classification problem with s groups, $\{G_1, G_2, ..., G_s\}$. One-Versus-All (OVA) is a technique that decomposes a problem with s groups into s binary subproblems where each problem is addressed by a binary classifier [21]. Each binary classifier, concerning the group u, is of the form:

$$G_u \text{ vs. } G_u^C \text{ , where } G_u^C = \{G_1, G_2, ..., G_s\} \setminus G_u$$

In this approach, all data are always considered in each binary classifier. The differentiation stands in the group assignment. The final assignment of which group a unit i belongs to is performed by using the majority vote.

3.3.1.3 Comparison

Among the advantages known, One-Versus-One (OVO) has a shorter training time than One-Versus-All (OVA), because the new binary problems are easier and smaller. In addition, One-Versus-One (OVO) usually obtains a higher performance [22]. Nevertheless, most of the binary subproblems created force data to be classified according to one of the two groups in the partition, even if it does not belong to either of the groups involved which may constitute a disadvantage.

Since in each binary subproblem all groups are tested, unlike One-Versus-One (OVO), One-Versus-All (OVA) does not force data to be classified between two groups. However, One-Versus-All (OVA) has a longer run time than One-Versus-One (OVO) and tends to create unbalanced data sets which may end up worsening the performance of the method.

Each binary subproblem created supplies a classification that is stored in order to be used for the final multi-class assignment. It is then reasonable to expect ties between groups. In the case of ties between groups with the highest number of votes, we conducted a random sample of size one between these tied groups.

Remark 3.7. Note that in each binary classification subproblem an optimisation step is required, since the matrices B and W change due to either group change or data change.

3.3.2 Consecutive Linear Discriminant Functions (CLDF)

3.3.2.1 Consecutive Linear Discriminant Functions (CLDF) for Classical Variables

For data with p explanatory variables and in order to discriminate s groups, at most δ linear discriminant functions may be defined, with $\delta = \min\{s-1, p\}$.

We start by recalling the classical case of multiple-group discriminant analysis explored in [23] and [24]. Consider p explanatory variables, $X_1, X_2 \ldots, X_p$. The first step is to estimate the first discriminant function. Suppose this function is

$$Z_1 = w_{11}X_1 + w_{12}X_2 + \ldots + w_{1p}X_p$$

where w_{ij} is the weight of the jth variable for the ith discriminant function. The weights of the first discriminant function are estimated by maximising:

$$\lambda_1 = \frac{\text{Between-groups Sums of Squares of } Z_1}{\text{Within-groups Sums of Squares of } Z_1}$$

Suppose that the second discriminant function is given by:

$$Z_2 = w_{21}X_1 + w_{22}X_2 + \ldots + w_{2p}X_p$$

The weights of the second discriminant function are estimated such that

$$\lambda_2 = \frac{\text{Between-groups Sums of Squares of } Z_2}{\text{Within-groups Sums of Squares of } Z_2}$$

is maximised subject to the constraint that the discriminant functions Z_1 and Z_2 are uncorrelated. This procedure is repeated until all possible discriminant functions are identified. This is an optimisation problem and the solution is given by the eigenvalues and eigenvectors of the non-symmetric matrix $W^{-1}B$, where W and B are the within-group and between-group Sums of Squares and Cross-Products (SSCP) matrices of the p variables, respectively. Since $W^{-1}B$ is non-symmetric, the eigenvectors may not be orthogonal, i.e., the discriminant functions may not be orthogonal, but they are uncorrelated.

3.3.2.2 Consecutive Linear Discriminant Functions (CLDF) for Histogram-valued Variables

Regarding Symbolic Data Analysis (SDA), this approach aims at defining several linear discriminant functions. Furthermore, these functions must be uncorrelated with each other. In other words, consider two optimal parameter vectors that define two distinct score histogram-valued variables. The symbolic linear correlation coefficient between the corresponding variables must be zero.

Consider a classification problem with s *a priori* classes, $\{G_1, G_2, ..., G_s\}$.

3.3.2.3 Optimisation

We start by addressing the optimisation step. Firstly, recall that, for data with s classes and p explanatory variables, at most $\delta = \min\{s-1, p\}$ functions are useful to discriminate between these classes. The optimisation step is performed iteratively. In each iteration, we are searching for the optimal parameter vector that defines a discriminant function. Moreover, apart from the first iteration, we must verify that the optimal parameter vector defines a new discriminant function that is not correlated with the discriminant functions previously found. Mathematically, at each time step u, for $u \in \{1, \ldots, \delta\}$, the optimal parameter vector γ_u is estimated according to the following optimisation formula:

$$\gamma_u^* = \arg\max_{\gamma_u} \lambda_u = \arg\max_{\gamma_u} \frac{\gamma_u^{\mathsf{T}} B \gamma_u}{\gamma_u^{\mathsf{T}} W \gamma_u}\,, \tag{3.11}$$

subject to:

$$\gamma_u \geq 0$$
$$\forall_{v \in \{1,\ldots,u-1\}} \quad r'(S^v, S^u) = 0$$

where S^v is the score's histogram-valued variable (also known as the linear discriminant function) that uses the optimal parameter vector γ_v, for $v \in \{1, \ldots, u-1\}$, obtained in previous iterations. S^u is the score's histogram-valued variable that uses the parameter vector γ_u, that we aim to optimise, for $u \in \{1, \ldots, \delta\}$.

3.3.2.4 Classification

Regarding the classification step, we introduce two possible definitions. Similarly to the previous classification definitions used, the goal is to identify the group that is at a minimum distance to a unit's score.

Definition 3.8. Consider the data divided in s *a priori* classes $\{G_1, \ldots, G_s\}$ and p explanatory variables, with $\delta = \min\{s-1, p\}$ linear discriminant functions that are useful to differentiate between these classes. Let $\Psi^v_{S(i)}$ be the score quantile function of unit i that uses the optimal parameter vector γ_v, for $v \in \{1, \ldots, \delta\}$. Moreover, let $\overline{\Psi^v_{S_u}}$ be the barycentric score associated with group u, for $u \in \{1, ..., s\}$, that uses the optimal parameter vector γ_v, for $v \in \{1, \ldots, \delta\}$.

For a given unit i, for $i \in \{1, ..., n\}$, the classification assignment can be given by the group G_u such that:

$$\min \sum_{v=1}^{\delta} D_M^2(\overline{\Psi_{S_u}^v}, \Psi_{S(i)}^v) \tag{3.12}$$

Definition 3.9. Consider the data divided in s *a priori* classes $\{G_1, \ldots, G_s\}$ and p explanatory variables, with $\delta = \min\{s-1, p\}$ linear discriminant functions that are useful to differentiate between these classes. Let $\Psi_{S(i)}^v$ be the score quantile function of unit i that uses the optimal parameter vector γ_v, for $v \in \{1, \ldots, \delta\}$. Using γ_v, let λ_v be the obtained ratio regarding the variability between groups and the variability within groups. Moreover, let $\overline{\Psi_{S_u}^v}$ be the barycentric score associated with group u, for $u \in \{1, ..., s\}$, that uses the parameter vector γ_v, for $v \in \{1, \ldots, \delta\}$.

For a given unit i, for $i \in \{1, ..., n\}$, the classification assignment can be given by the group G_u such that:

$$\min \sum_{v=1}^{\delta} \left(\frac{\lambda_v}{\sum\limits_{j=1}^{\delta} \lambda_j} D_M^2(\overline{\Psi_{S_u}^v}, \Psi_{S(i)}^v) \right) \tag{3.13}$$

Once more it is reasonable to account for the cases where there are units equally distanced from the barycentric scores of several groups, with the minimum distance. In case there is more than one group that satisfies the condition 3.12 or condition 3.13, according to the definition used, the group assigned is the result of a random sample of size one over the set containing those groups, without replacement. Note that this procedure may introduce randomness to the classification process.

The Definition 3.8 aims at identifying the group that minimises the overall distance to the unit's score. For every discriminant function, the distance between the barycentric score and the unit's score is considered. On the other hand, the Definition 3.9 aims at identifying the group that is closest to a given unit i, taking into account how well each discriminant function separates the classes. In order to do that, we perform a weighted average of the Mallows distance between the unit's score and the barycentric score of each group that is defined by the *a priori* classes. The weights convey the importance a specific discriminant function has in separating the data into classes. The overall separation of the data in groups is given by the summation of all λ_j, for $j \in \{1, \ldots, \delta\}$, since all discriminant functions obtained are useful and uncorrelated. Moreover, a specific λ_v measures how well the discriminant function that uses the optimal parameter vector γ_v separates data into groups. Therefore, the weights are given by the ratio between the measure associated with how well a specific discriminant function separates the data into groups and the measure associated with how well all discriminant functions separate the data into groups together.

3.3.2.5 Comparison

Like One-Versus-All (OVA), Consecutive Linear Discriminant Functions (CLDF) always considers the entire data set. Contrary to what happens with both One-Versus-One (OVO) and One-Versus-All (OVA), this method does not create binary subproblems. This translates into a computational advantage, i.e., the run time is fairly lower.

Chapter 4

Implementation

In this chapter, we present the general concepts of the implementation in R of the linear discriminant analysis method developed in [1] and the proposed extension in this thesis. We start by addressing the data structures of symbolic data used and the main functions developed. Lastly, it is presented the optimisation implementation idea as well as decisions regarding the classification process.

4.1 R packages

The code developed requires the loading of the following R packages: *caret*, *ggplot*2, *HistDAWass* [25], *Rcpp*, *Rcsdp* [26], *ROI* and *ROI.plugin.alabama* [27].

4.2 Data Structure

The data structure used in this thesis takes into consideration the framework of Symbolic Data Analysis (SDA) used in the package developed by Professor Antonio Irpino, *HistDAWass* [25]. This package introduces *distributionH* object and *MatH* (matrix of distributions) object. Regarding the functions that belong to this package, these objects were by far the most used in this implementation.

4.2.1 Class *distributionH*

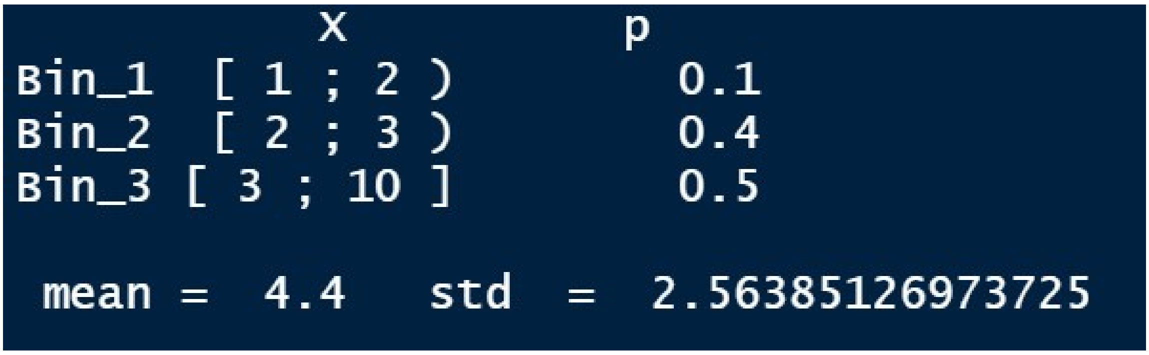
The function *distributionH* creates a histogram object. This object contains the description of a histogram that corresponds to an entry or a cell in a symbolic data set. The arguments used are x which is a numeric vector that is the domain of the distribution (extremes of bins) and p a non-decreasing and non-negative numeric vector with the same length as x which is the cumulative distribution function.

Example 4.1. *Consider the following code:*

```
distributionH(x = c(1, 2, 3, 10), p = c(0, 0.1, 0.5, 1))
```

LISTING 4.1: Creation of a *distributionH* object

The R console output is the following:



```
              X           p
Bin_1  [ 1 ; 2 )         0.1
Bin_2  [ 2 ; 3 )         0.4
Bin_3 [ 3 ; 10 ]         0.5

 mean =  4.4   std  =  2.56385126973725
```

FIGURE 4.1: Example of *distributionH* object output in R console.

4.2.2 Class *MatH*

The class *MatH* defines a matrix of *distributionH* objects. This function creates a symbolic data set. It generates a matrix of histogram data, in other words, a *MatH* object.

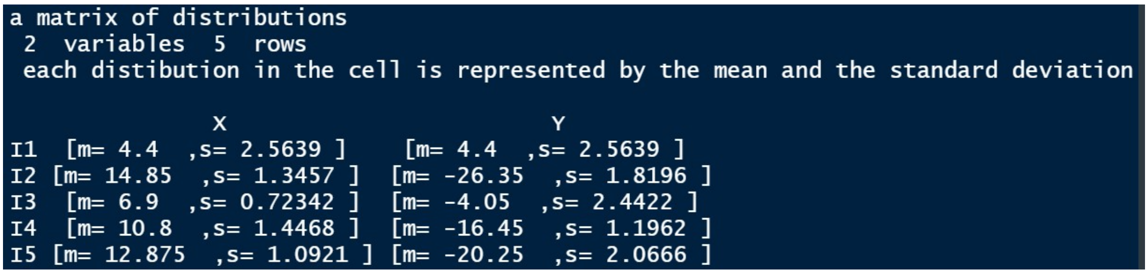
Example 4.2. *Consider the following code:*

```
a1 <- distributionH(x = c(1, 2, 3, 10), p = c(0, 0.1, 0.5, 1))
a2 <- distributionH(x=c(12,13,15,17), p = c(0, 0.1, 0.5, 1))
a3 <- distributionH(x=c(5,6,7,8), p = c(0, 0.1, 0.5, 1))
a4 <- distributionH(x=c(7,9,11,13), p = c(0, 0.1, 0.5, 1))
a5 <- distributionH(x=c(10,12,12.5, 15), p = c(0, 0.1, 0.5, 1))
b1 <- a1
b2 <- distributionH(x=c(-33,-28,-26,-24), p = c(0, 0.1, 0.5, 1))
b3 <- distributionH(x=c(-10,-7,-4,0), p = c(0, 0.1, 0.5, 1))
b4 <- distributionH(x=c(-20,-18,-16,-15), p = c(0, 0.1, 0.5, 1))
b5 <- distributionH(x=c(-25,-23,-20,-17), p = c(0, 0.1, 0.5, 1))


a <- MatH(x=c(a1, a2, a3, a4, a5, b1, b2, b3, b4, b5),
          nrows =5, ncols = 2, varnames = c("X", "Y"))
```

LISTING 4.2: Creation of a *MatH* object

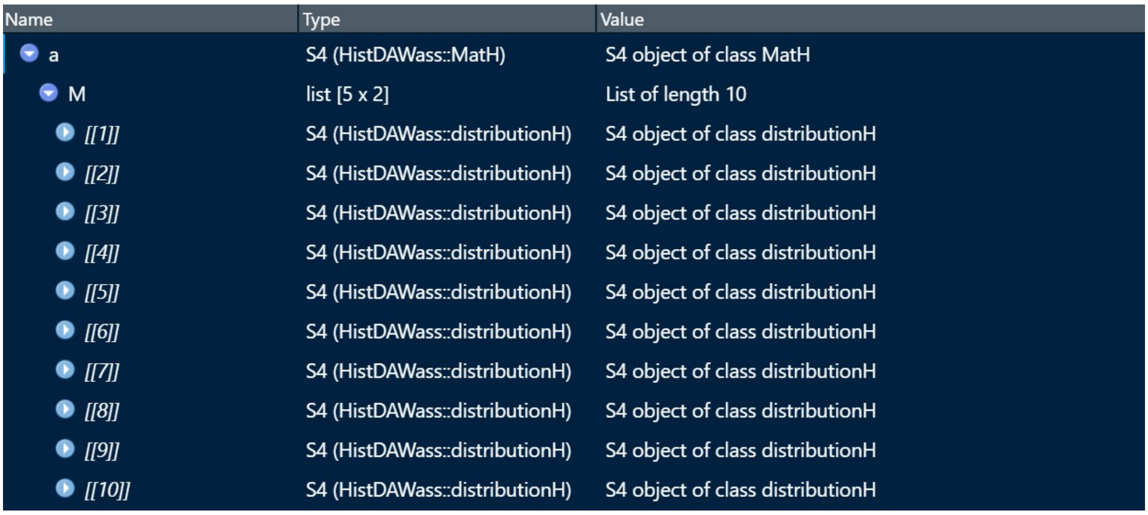
It created a symbolic data table with 5 units described by 2 histogram-valued variables, X and Y. The R output is the following:



```
a matrix of distributions
 2  variables  5  rows
 each distibution in the cell is represented by the mean and the standard deviation

                X                         Y
I1  [m= 4.4  ,s= 2.5639 ]    [m= 4.4  ,s= 2.5639 ]
I2 [m= 14.85  ,s= 1.3457 ]  [m= -26.35  ,s= 1.8196 ]
I3  [m= 6.9  ,s= 0.72342 ]  [m= -4.05  ,s= 2.4422 ]
I4  [m= 10.8  ,s= 1.4468 ]  [m= -16.45  ,s= 1.1962 ]
I5 [m= 12.875  ,s= 1.0921 ] [m= -20.25  ,s= 2.0666 ]
```

FIGURE 4.2: Example of *MatH* object output in R console.



Name	Type	Value
a	S4 (HistDAWass::MatH)	S4 object of class MatH
M	list [5 x 2]	List of length 10
[[1]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[2]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[3]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[4]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[5]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[6]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[7]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[8]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[9]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH
[[10]]	S4 (HistDAWass::distributionH)	S4 object of class distributionH

FIGURE 4.3: Example of formal class *MatH* data environment.

4.3 Functions developed

In this section, we introduce the main functions developed to allow the application of Linear Discriminant Analysis. Note that these functions may rely on several others of minor importance in order to achieve the desired purpose.

- *getSquaredMallowsDistance*: given two *distributionH* objects, this function calculates the squared Mallows distance according to the rewritten formula 2.12.

- *RewrittingOperation*: given a *MatH* object, this function performs the rewriting operation described in Chapter 2.

- *B_matrix* and *W_matrix*: given the data set, as a *MatH* object, and a numeric vector containing the ground truth regarding the *a priori* classes, these functions compute the matrices of the SSCP between-groups and within-groups, respectively.

- *Lambdaoptimisation*: given the data set as a *MatH* object and a numeric vector containing the ground truth regarding the *a priori* classes, this function performs the optimisation processes needed in order to find the optimal parameter vector. The output of this function is either:

  - only the parameter vector, if it is a binary classification problem
  - or a list of objects containing the parameter vectors, the λ values associated and the summation of all the λ values, in case it is a multi-class classification problem.

- *LDA2Class*: this function performs the classification into two *a priori* groups. The inputs are both the training and testing data sets and also the numeric vector containing the ground truth regarding the *a priori* classes. The optimisation process is performed inside this function.

- *LDAClassification*: this function performs the multi-class classification, considering as an input, among others, the classification technique desired: One-Versus-All (OVA), One-Versus-One (OVO) or Consecutive Linear Discriminant Functions (CLDF).

- *TrainTestSplit*: since the data structure is not a typical one, the splitting between training and testing data sets may be difficult. Therefore, this function aims to help when in this situation. The output is a list with 4 objects: the training data set, the ground truth regarding classes associated with the training data set, the testing data set and the ground truth regarding classes associated with the testing data set.

4.4 Optimisation

The constrained quadratic problem compelled us to operate with two optimisation algorithms, in order to prove optimality, already implemented in RStudio:

1. One that provides an admissible solution, since the algorithm does not converge.

2. Another one that is obtained according to the semi-definite programming relaxations formulation that aims at finding a good upper bound to this problem.

When used together, it is proven computationally that the solution obtained is optimal [18].

The optimisation algorithm used to find an admissible solution, in RStudio, uses the function *ROI_solve* that requires the library *ROI*. The *solver* used was *alabama* [27], when the correlation restriction is used, and *nlminb* in the other cases. The optimisation algorithm used to run the semi-definite programming relaxations uses the function *csdp* that requires the library *Rcsdp* [26].

The algorithms used return a message parameter that gives out information regarding the feasibility of both the primal and dual problems. These parameters were taken into account when developing the code regarding the optimisation required for this problem.

This process introduces problems that are typical in the optimisation field and may be encountered. For example, when the program is stuck at an edge of feasibility or even infeasibility. In these cases, we are unable to obtain an optimal parameter vector.

4.5 Classification

As previously mentioned, the class assignment is based on distances to the barycentric score of the classes and/or based on the majority vote. Theoretically, the *sample* operation is only performed when a given unit is equally distanced from the barycentric scores of different classes. In practice, due to rounding errors and the machine's precision, we defined a threshold, called *epsilon*, of 10^{-5} that determines whether the distances are close enough to be considered equal.

The code developed keeps track of the *sample* operations performed both in binary and multi-class classifications. Not only it counts the number of *sample* operations, but also displays in which units it was performed.

4.5.1 Two *A Priori* Groups

Using the Mallows distance, for each unit, we calculate the distance between the score quantile function associated with the given unit and the barycentric scores of all the groups, assigning the unit to the group whose barycentric score is closest to the unit's score.

The implementation of this technique only relies on the construction of the vector with the groups assigned.

4.5.2 One-Versus-All (OVA)

The implementation of this technique uses a matrix with as many lines as the number of classes and as many columns as the units in the testing data set. Each binary subproblem provides a line in that matrix. The classification that comes from the binary subproblem G_u vs. $G_u^C = \{G_1, G_2, ..., G_s\} \setminus G_u$ is a vector in which an entry is 1 when the binary classifier classifies a given unit with G_u and 0, otherwise.

The classification vectors obtained from every binary classifier associated with the subproblems are then used for the multi-class classification using the majority vote per unit. To clarify, the binary subproblems fill the matrix. Then, for each unit, we look to the column associated and search for the group that has the highest number of votes. If there is more than one group with the maximum number of votes, we perform a sample operation of size one over the set containing these groups.

4.5.3 One-Versus-One (OVO)

The implementation of this technique also uses a matrix with as many lines as the number of classes and as many columns as the units in the testing data set. Each binary subproblem provides a vote per unit, considering the classes tested in the subproblem. This matrix is then used for the multi-class classification using the majority vote per unit.

4.5.4 Consecutive Linear Discriminant Functions (CLDF)

The implementation of the classification of this method is somewhat similar to the binary classification. For each unit, the classification assignment is given by the group identified, following the Definitions 3.8 or 3.9.

The implementation of this technique only relies on the construction of the vector with the groups assigned.

Remark 4.1. Symbolic data with a considerable amount of variables, units and/or weights severely impact the run time of the Linear Discriminant Analysis. Moreover, having a large number of subintervals, therefore a large number of weights, causes the distribution of the data to be meaningless. In order to avoid this situation, we may consider histograms with equal probability subintervals (equiprobable histograms), for every unit [28].

Example 4.3. *This example aims at clarifying how the multi-class classification step is performed, especially for the strategies One-Versus-All (OVA) and One-Versus-One (OVO), since the classification of Consecutive Linear Discriminant Functions (CLDF) relies only on the calculation of the Mallows distances.*

Consider a multi-class classification problem with 3 classes where the classification assignment is given in Table 4.1.

Unit	1	2	3	4	5	6	7	8
Class	1	2	3	3	1	1	2	1

TABLE 4.1: Classification assignment.

One-Versus-All (OVA): *We start with the matrix in Table 4.2 with as many columns as the number of units and as many lines as the number of classes.*

		Unit							
		1	2	3	4	5	6	7	8
Pred.	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0

TABLE 4.2: Initial predicted classification matrix for One-Versus-All (OVA).

Recall that each binary subproblem provides a line in this matrix, therefore, consider the results displayed in Tables 4.3, 4.4 and 4.5.

Unit	1	2	3	4	5	6	7	8
Pred	0	0	1	0	1	1	0	1

(A) Predicted classification for the binary subproblem G_1 vs. G_1^C.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	0	0	1	0	1	1	0	1
	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0

(B) Predicted classification matrix updated.

TABLE 4.3: Results after the binary subproblem G_1 vs. G_1^C in One-Versus-All (OVA).

Unit	1	2	3	4	5	6	7	8
Pred	0	1	0	0	1	0	1	1

(A) Predicted classification for the binary subproblem G_2 vs. G_2^C.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	0	0	1	0	1	1	0	1
	2	0	1	0	0	1	0	1	1
	3	0	0	0	0	0	0	0	0

(B) Predicted classification matrix updated.

TABLE 4.4: Results after the binary subproblem G_2 vs. G_2^C in One-Versus-All (OVA).

Unit	1	2	3	4	5	6	7	8
Pred	0	0	0	1	0	1	1	1

(A) Predicted classification for the binary subproblem G_3 vs. G_3^C.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	0	0	1	0	1	1	0	1
	2	0	1	0	0	1	0	1	1
	3	0	0	0	1	0	1	1	1

(B) Predicted classification matrix updated.

TABLE 4.5: Results after the binary subproblem G_3 vs. G_3^C in One-Versus-All (OVA).

The next step is to analyse Table 4.5b in order to determine which groups have the maximum number of votes per unit. Note that units 1, 5, 6, 7 and 8 have several groups in this situation, therefore we must perform sample operations. Suppose the results of the sample operations are the ones given below:

- *Unit 1: sample(1,2,3) = 2;*
- *Unit 5: sample(1,2) = 2;*
- *Unit 6: sample(1,3) = 1;*
- *Unit 7: sample(2,3) = 3;*
- *Unit 8: sample(1,2,3) = 2;*

The classification given by One-Versus-All (OVA) is:

Unit	1	2	3	4	5	6	7	8
Pred	2	2	1	3	2	1	3	2

TABLE 4.6: Predicted classification by One-Versus-All (OVA).

***One-Versus-One (OVO):** We start with the matrix in Table 4.7 with as many columns as the number of units and as many lines as the number of classes.*

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0

TABLE 4.7: Initial predicted classification matrix for One-Versus-One (OVO).

Recall that each binary subproblem provides a vote per unit, considering the classes tested in the subproblem. Consider the results displayed in Tables 4.8, 4.9 and 4.10.

Unit	1	2	3	4	5	6	7	8
Pred	1	2	1	2	1	2	2	1

(A) Predicted classification for the binary subproblem G_1 vs. G_2.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	1	0	1	0	1	0	0	1
	2	0	1	0	1	0	1	1	0
	3	0	0	0	0	0	0	0	0

(B) Predicted classification matrix updated.

TABLE 4.8: Results after the binary subproblem G_1 vs. G_2 in One-Versus-One (OVO).

Unit	1	2	3	4	5	6	7	8
Pred	2	2	3	3	2	3	2	3

(A) Predicted classification for the binary subproblem G_2 vs. G_3.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	1	0	1	0	1	0	0	1
	2	1	2	0	1	1	1	2	0
	3	0	0	1	1	0	1	0	1

(B) Predicted classification matrix updated.

TABLE 4.9: Results after the binary subproblem G_2 vs. G_3 in One-Versus-One (OVO).

Unit	1	2	3	4	5	6	7	8
Pred	1	1	3	3	3	1	3	1

(A) Predicted classification for the binary subproblem G_1 vs. G_3.

		Unit							
		1	2	3	4	5	6	7	8
Pred	1	2	1	1	0	1	1	0	2
	2	1	2	0	1	1	1	2	0
	3	0	0	2	2	1	1	1	1

(B) Predicted classification matrix updated.

TABLE 4.10: Results after the binary subproblem G_1 vs. G_3 in One-Versus-One (OVO).

The next step is to analyse Table 4.10b in order to determine which groups have the maximum number of votes per unit. Note that units 5 and 6 have several groups in this situation, therefore we must perform sample operations. Suppose the results of the sample operations are the ones given below:

- *Unit 5: sample(1,2,3) = 1;*
- *Unit 6: sample(1,2,3) = 2;*

The classification given by One-Versus-One (OVO) is:

Unit	1	2	3	4	5	6	7	8
Pred	1	2	3	3	1	2	2	1

TABLE 4.11: Predicted classification by One-Versus-One (OVO).

Chapter 5

Application

In this section, we present several applications of the model developed, considering the three techniques explored. Several network data sets that were artificially created are analysed. In this thesis, networks are referenced in the sense of graphs. It is given a brief introduction to the relevant concepts in respect of network science. Moreover, we studied the model's performance on four data sets regarding Internet traffic attacks. These attacks are the result of the corruption of the Border Gateway Protocol (BGP), resulting in redirecting [29]. Finally, we studied the techniques' performance on interval-valued variables. This is explored by using a small data set associated with car models.

5.1 Performance Measures

In classification problems, in order to assess a model's performance, it is important to choose measures that reflect the ability that the model has to predict the correct classes. In this sense, the measures used were accuracy and balanced accuracy.

The accuracy formula is the following:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Number of Predictions}}$$

Balanced accuracy is employed when evaluating the strategy's predictive ability for each class separately. The main advantage is that this measure accounts for imbalanced classes. The formula for a binary classification is the following:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{Correct Positive Predictions}}{\text{Number of Positives}} + \frac{\text{Correct Negative Predictions}}{\text{Number of Negative}}\right)$$

5.2 Modelling Network Data

5.2.1 Description of the Data

The network data sets are synthetic. We developed networks considering the Erdős-Renyi, Watts–Strogatz and Barabási–Albert models, with the parameters carefully chosen. The classification problem relies on identifying the model used to develop each network (3 classes):

- 1 - Barabási–Albert;
- 2 - Erdős-Renyi;
- 3 - Watts–Strogatz;

5.2.2 Network Models

In order to better understand the following information, it is suggested that the reader gets acquainted with concepts associated with network science such as degree, betweenness centrality, closeness centrality, eigenvector centrality, giant component and regular graph that can be found in [30].

5.2.2.1 Erdős-Renyi

This is the simplest model in network science. The generation of a random network following this model consists of N nodes where each pair of nodes is connected with a probability p. The degree (number of neighbours) distribution of a random network follows a Binomial distribution, $Binomial(N, p)$, therefore the average degree, $< k >$, is given by:

$$< k >= p(N-1) \tag{5.1}$$

For different values of p, we have different regimes. For $p = 0$, all nodes are isolated and, for $p = 1$, we have a graph that is fully connected. Figure 5.1, which can be found in [31] and [30], displays the different regimes that an Erdős-Renyi network can be in.

The values used to produce the Erdős-Renyi networks in the data sets were:

- $N \in \{100, 200, 300, 400, 500, 550, 600, 700\}$;
- $p \in \{0.01, 0.03, 0.05, 0.1, 0.15, 0.2, 0.25\}$.

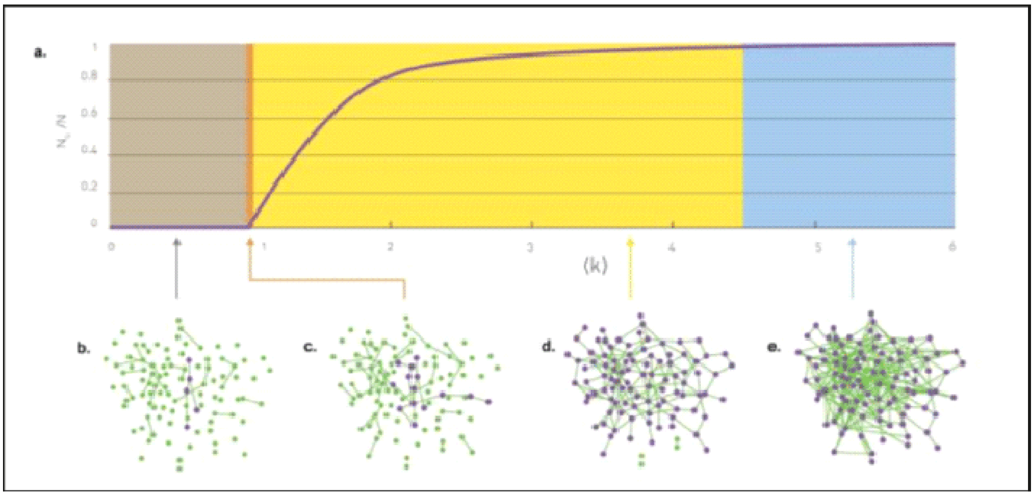


FIGURE 5.1: (a) Number of nodes in the giant component (normalized) in the Erdős-Renyi network plotted as a function of the average degree. (b) Subcritical regime: few small clusters are present. (c) Phase transition at the critical point of $< k >= 1$. (d) Supercritical regime: a larger fraction of nodes belong to the giant component. (e) Connected regime: the giant component contains almost every node in the network. Image source: [30].

5.2.2.2 Watts–Strogatz

This model was developed in order to merge the network's structure and randomness. This model only has one parameter that controls the probability of rewiring an edge, p. Starting from a regular graph (a graph where each vertex has the same number of neighbours), each edge is randomly redirected with a chosen probability. Figure 5.2 that can be found in [32] aims at displaying the network effect when changing the rewiring probability.

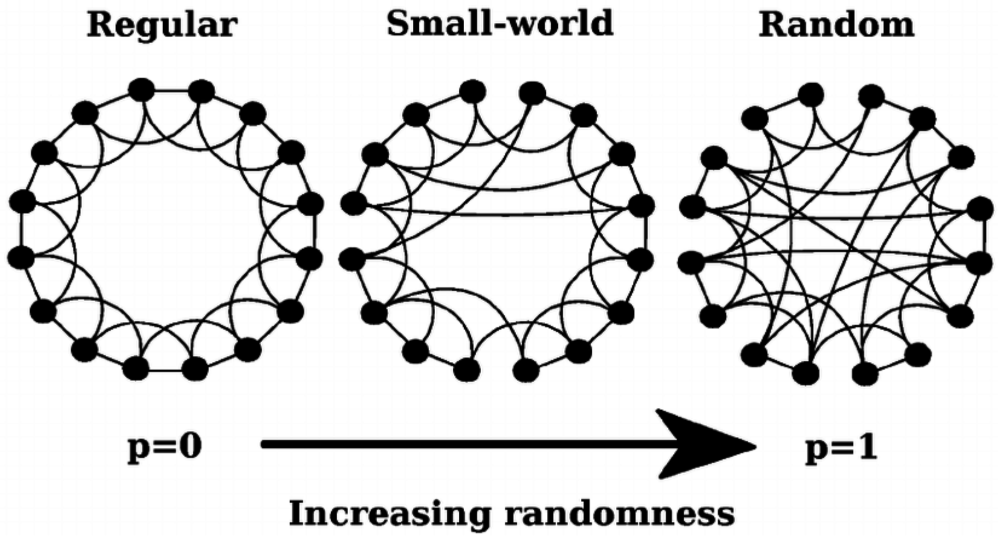


FIGURE 5.2: Algorithm that produces Watts-Strogatz networks, according to the rewiring probability p.

The networks produced when $0 < p < 1$ display a mixture of regular and random connections.

The values used to produce the Watts–Strogatz networks in the data sets were:

- $N \in \{100, 200, 300, 400, 500, 550, 600, 700\}$;
- $p \in \{0.01, 0.05, 0.2, 0.4, 0.5, 0.6, 0.8\}$.

5.2.2.3 Barabási–Albert

This model generates networks known as scale-free networks, where the degree distributions decay with a power law. This model is based on both growth and preferential attachment. The first concept is obtained by adding nodes sequentially. The second one is obtained by taking into account the already existing nodes' degree when establishing a new connection with a new node added to the network. It favours new connections with already highly connected nodes. An analogy commonly referred to is *"the more popular you are, the more popular you become"*.

Consider a network with N final nodes and $m0$ initial nodes. Sequentially, we add one node and create m new edges with the new node and already existing ones. After a fair amount of time steps, t, the graph has $m_0 + t$ nodes and mt edges, therefore $< k >\sim 2m$.

The values used to produce the Barabási–Albert networks in the data sets were:

- $N \in \{100, 200, 300, 400, 500, 550, 600, 700\}$;
- $(m0, m) \in \{(3, 1), (5, 2), (7, 3), (9, 4), (11, 5), (13, 6), (15, 7)\}$.

5.2.3 Data sets

Each network (unit in the symbolic data sets) is described by the distribution over the network's nodes of standard graph centrality measures:

- Degree: number of neighbours;
- Betweenness centrality;
- Closeness centrality;
- Eigenvector centrality;

Therefore, the dataset has four histogram-valued variables.

In view of testing the performance of the approaches developed on both balanced and unbalanced data sets, we developed eight different network data sets. Each dataset has a fixed number of units per network model. The description of each dataset is given in Table 5.1.

	Erdős-Renyi	Watts–Strogatz	Barabási–Albert	Total
Balanced	56	56	56	168
UnbalER	12	56	56	124
UnbalWS	56	12	56	124
UnbalBA	56	56	12	124
UnbalERWS	12	12	56	80
UnbalWSBA	56	12	12	80
UnbalERBA	12	56	12	80
UnbalAll	54	28	18	100

TABLE 5.1: Number of units per network model in each synthetic dataset.

Note that each standard graph measure is associated with a node, i.e., the degree, betweenness centrality, closeness centrality and eigenvector centrality are measured with respect to a node in a network. Therefore, the first-level units are the nodes. To obtain symbolic data sets, we performed contemporary aggregations, considering all the nodes that belong to a specific network. Therefore, the higher-level units are the networks. For example, in the symbolic data sets, the concretisations of the histogram-valued variable Degree can be seen as the degree distribution of each network produced.

5.2.4 Discussion of Results

We divided each data set into training and testing data sets considering 80% and 20%, respectively. The following tables display the confusion matrices obtained from applying the One-Versus-All (OVA), One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) strategies to the several networks data sets created. ”CLDF w/ weights” references the Consecutive Linear Discriminant Functions (CLDF) strategy according to the Definition 3.9 and ”CLDF w/o weights” references the Consecutive Linear Discriminant Functions (CLDF) strategy according to the Definition 3.8. Moreover, in the tables that display the number of sample operations performed ”MV” references the samples performed in the majority vote for the multi-class classification strategies.

		Reference		
		1	2	3
Predicted	1	35	2	0
	2	10	17	24
	3	1	24	20

(A) OVA.

		Reference		
		1	2	3
Predicted	1	46	0	0
	2	0	43	5
	3	0	0	39

(B) OVO.

		Reference		
		1	2	3
Predicted	1	45	1	0
	2	1	42	15
	3	0	0	29

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	46	1	0
	2	0	42	11
	3	0	0	33

(D) CLDF w/o weights.

TABLE 5.2: Confusion matrices obtained from applying the strategies to the training data set of *Balanced*.

		Reference		
		1	2	3
Predicted	1	2	3	4
	2	7	4	4
	3	1	6	4

(A) OVA.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	1	13	1
	3	0	0	11

(B) OVO.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	1	9	0
	3	0	4	12

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	1	11	2
	3	0	2	10

(D) CLDF w/o weights.

TABLE 5.3: Confusion matrices obtained from applying the strategies to the testing data set of *Balanced*.

		Reference		
		1	2	3
Predicted	1	46	2	0
	2	0	4	4
	3	0	5	39

(A) OVA.

		Reference		
		1	2	3
Predicted	1	46	0	0
	2	0	11	2
	3	0	0	41

(B) OVO.

		Reference		
		1	2	3
Predicted	1	45	1	0
	2	1	10	19
	3	0	0	24

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	46	2	0
	2	0	9	15
	3	0	0	28

(D) CLDF w/o weights.

TABLE 5.4: Confusion matrices obtained from applying the strategies to the training data set of *UnbalER*.

		Reference		
		1	2	3
Predicted	1	3	0	5
	2	4	1	2
	3	3	0	6

(A) OVA.

		Reference		
		1	2	3
Predicted	1	10	1	0
	2	0	0	1
	3	0	0	12

(B) OVO.

		Reference		
		1	2	3
Predicted	1	4	0	0
	2	6	1	0
	3	0	0	13

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	7	1	1
	3	0	0	12

(D) CLDF w/o weights.

TABLE 5.5: Confusion matrices obtained from applying the strategies to the testing data set of *UnbalER*.

		Reference		
		1	2	3
Predicted	1	44	1	0
	2	0	15	4
	3	1	28	6

(A) OVA.

		Reference		
		1	2	3
Predicted	1	43	0	0
	2	2	31	0
	3	0	13	10

(B) OVO.

		Reference		
		1	2	3
Predicted	1	40	0	0
	2	5	40	1
	3	0	4	9

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	45	1	0
	2	0	43	6
	3	0	0	4

(D) CLDF w/o weights.

TABLE 5.6: Confusion matrices obtained from applying the strategies to the training data set of *UnbalWS*.

		Reference		
		1	2	3
Predicted	1	4	1	1
	2	6	8	1
	3	1	3	0

(A) OVA.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	1	8	0
	3	0	4	2

(B) OVO.

		Reference		
		1	2	3
Predicted	1	11	0	0
	2	0	11	0
	3	0	1	2

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	1	12	0
	3	0	0	2

(D) CLDF w/o weights.

TABLE 5.7: Confusion matrices obtained from applying the strategies to the testing data set of *UnbalWS*.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	0	32	27
	3	0	16	14

(A) OVA.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	0	48	6
	3	0	0	35

(B) OVO.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	0	48	11
	3	0	0	30

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	10	0	0
	2	0	48	13
	3	0	0	28

(D) CLDF w/o weights.

TABLE 5.8: Confusion matrices obtained from applying the strategies to the training data set of *UnbalBA*.

		Reference		
		1	2	3
Predicted	1	1	3	4
	2	1	2	6
	3	0	3	5

(A) OVA.

		Reference		
		1	2	3
Predicted	1	2	0	0
	2	0	8	1
	3	0	0	14

(B) OVO.

		Reference		
		1	2	3
Predicted	1	1	0	0
	2	1	8	1
	3	0	0	14

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	2	0	0
	2	0	8	0
	3	0	0	15

(D) CLDF w/o weights.

TABLE 5.9: Confusion matrices obtained from applying the strategies to the testing data set of *UnbalBA*.

		Reference		
		1	2	3
Predicted	1	43	2	0
	2	0	3	0
	3	0	6	10

(A) OVA.

		Reference		
		1	2	3
Predicted	1	43	0	0
	2	0	11	1
	3	0	0	9

(B) OVO.

		Reference		
		1	2	3
Predicted	1	42	1	0
	2	1	10	5
	3	0	0	5

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	42	2	0
	2	1	9	4
	3	0	0	6

(D) CLDF w/o weights.

TABLE 5.10: Confusion matrices obtained from applying the strategies to the training data set of *UnbalERWS*.

		Reference		
		1	2	3
Predicted	1	6	0	0
	2	2	1	1
	3	5	0	1

(A) OVA.

		Reference		
		1	2	3
Predicted	1	11	0	0
	2	2	1	0
	3	0	0	2

(B) OVO.

		Reference		
		1	2	3
Predicted	1	8	0	0
	2	5	1	0
	3	0	0	2

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	12	0	0
	2	1	1	0
	3	0	0	2

(D) CLDF w/o weights.

TABLE 5.11: Confusion matrices obtained from applying the strategies to the testing data set of $UnbalERWS$.

		Reference		
		1	2	3
Predicted	1	6	0	0
	2	2	14	3
	3	1	29	5

(A) OVA.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	0	42	1
	3	0	1	7

(B) OVO.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	0	43	3
	3	0	0	5

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	7	0	0
	2	2	31	0
	3	0	12	8

(D) CLDF w/o weights.

TABLE 5.12: Confusion matrices obtained from applying the strategies to the training data set of $UnbalWSBA$.

		Reference		
		1	2	3
Predicted	1	0	5	1
	2	1	6	1
	3	2	2	2

(A) OVA.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	0	13	1
	3	0	0	3

(B) OVO.

		Reference		
		1	2	3
Predicted	1	3	1	0
	2	0	12	0
	3	0	0	4

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	0	11	1
	3	0	2	3

(D) CLDF w/o weights.

TABLE 5.13: Confusion matrices obtained from applying the strategies to the testing data set of $UnbalWSBA$.

		Reference		
		1	2	3
Predicted	1	9	1	0
	2	0	2	4
	3	0	8	40

(A) OVA.

		Reference		
		1	2	3
Predicted	1	9	0	0
	2	0	11	1
	3	0	0	43

(B) OVO.

		Reference		
		1	2	3
Predicted	1	8	0	0
	2	1	11	5
	3	0	0	39

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	8	0	0
	2	1	11	7
	3	0	0	37

(D) CLDF w/o weights.

TABLE 5.14: Confusion matrices obtained from applying the strategies to the training data set of *UnbalERBA*.

		Reference		
		1	2	3
Predicted	1	1	0	5
	2	2	0	3
	3	0	1	4

(A) OVA.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	0	1	2
	3	0	0	10

(B) OVO.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	0	1	4
	3	0	0	8

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	3	0	0
	2	0	1	0
	3	0	0	12

(D) CLDF w/o weights.

TABLE 5.15: Confusion matrices obtained from applying the strategies to the testing data set of *UnbalERBA*.

		Reference		
		1	2	3
Predicted	1	15	0	0
	2	0	25	11
	3	1	20	9

(A) OVA.

		Reference		
		1	2	3
Predicted	1	16	0	3
	2	0	44	1
	3	0	1	16

(B) OVO.

		Reference		
		1	2	3
Predicted	1	16	0	0
	2	0	41	7
	3	0	4	13

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	16	0	0
	2	0	45	8
	3	0	0	12

(D) CLDF w/o weights.

TABLE 5.16: Confusion matrices obtained from applying the strategies to the training data set of *UnbalAll*.

		Reference		
		1	2	3
Predicted	1	1	4	1
	2	1	0	5
	3	0	5	2

(A) OVA.

		Reference		
		1	2	3
Predicted	1	2	0	1
	2	0	9	0
	3	0	0	7

(B) OVO.

		Reference		
		1	2	3
Predicted	1	1	0	0
	2	1	9	0
	3	0	0	8

(C) CLDF w/ weights.

		Reference		
		1	2	3
Predicted	1	2	0	0
	2	0	9	1
	3	0	0	7

(D) CLDF w/o weights.

TABLE 5.17: Confusion matrices obtained from applying the strategies to the testing data set of *UnbalAll*.

By observing the confusion matrices in Tables 5.2 to 5.17, we observe that there is rarely any confusion in distinguishing classes 1 and 3. The most frequent confusion between classes concerns classes 2 and 3. Occasionally, we observed confusion between classes 1 and 2.

Moreover, when unbalancing the classes, it is possible to realise that the error rate may be higher for the minority classes. For example, in Table 5.4a, the minority class is 2. Regarding the number of units that belong to class 2, we verify that approximately 63.64% (7/11) were classified as belonging to the other classes. Regarding class 1, 0% were classified as belonging to the other classes and, regarding class 3, 9.30% (4/43) were classified as belonging to the class 2. Therefore, the error rate associated with class 2 is higher than the ones associated with the remaining classes.

It is possible to understand that the One-Versus-All (OVA) strategy displays the worst performance. Notice that One-Versus-One (OVO) performs extremely well, regardless of the model used to produce the networks. Moreover, Consecutive Linear Discriminant Functions (CLDF) strategies also perform fairly well.

Table 5.18 considers the global accuracy associated with the classifications obtained by applying each strategy developed. As expected, the accuracy values associated with the training data set tend to be higher than those associated with the testing data set.

In the vast majority of cases, the One-Versus-All (OVA) strategy does not provide good models to classify the data. This is clearly evident when analysing the testing data set results. Even with a fairly good fit on the training data set, the generalisation is not adequate. For previously unseen data, the accuracy values registered can reach 50.00% or

	OVA		OVO		CLDF w/ weights		CLDF w/o weights	
	train	test	train	test	train	test	train	test
Balanced	54.14	28.57	96.24	94.29	87.22	85.71	90.98	85.71
UnbalER	89.00	41.67	98.00	91.67	79.00	75.00	83.00	66.67
UnbalWS	65.66	48.00	84.85	80.00	89.90	96.00	92.93	96.00
UnbalBA	56.57	32.00	93.94	96.00	88.89	92.00	86.87	100.00
UnbalERWS	87.50	50.00	98.44	87.50	89.06	68.75	89.06	93.75
UnbalWSBA	41.67	40.00	96.67	95.00	95.00	95.00	76.67	85.00
UnbalERBA	79.69	31.25	98.44	87.50	90.62	75.00	87.50	100.00
UnbalAll	60.49	15.79	93.83	94.74	86.42	94.74	90.12	94.74

TABLE 5.18: Accuracy for the multi-class classification strategies (%).

can be as low as 15.79% which is unacceptable.

On the other hand, One-Versus-One (OVO) and both Consecutive Linear Discriminant Functions (CLDF) strategies tend to provide good models. One-Versus-One (OVO) performs extremely well, registering accuracy values consistently higher than 84% on the training data set and higher than 80% on the testing data set. In respect of the *UnbalAll*, where the accuracy of the testing data set is higher than the accuracy of the training data set, it is possible that the model is incurring in underfitting in the optimisation step. This also tends to happen when using the strategies of Consecutive Linear Discriminant Functions (CLDF) according to the Definition 3.8, i.e., without using weights. Although One-Versus-One (OVO) performs better than Consecutive Linear Discriminant Functions (CLDF), these strategies tend to properly fit the data. Regarding the Consecutive Linear Discriminant Functions (CLDF), when considering the Definition 3.9, i.e., using weights, the worst model registered has associated accuracy values of 79% and 75% on the training and testing data sets, respectively, still providing a good fit. Regarding the Consecutive Linear Discriminant Functions (CLDF), when considering the Definition 3.8, i.e., without using weights, the worst model registered has accuracy values of 83% on the training and 66.67% on the testing data sets.

Therefore, One-Versus-One (OVO) and both Consecutive Linear Discriminant Functions (CLDF) strategies tend to perform well. Moreover, this performance is better than the one associated with One-Versus-All (OVA). It is possible to realise that unbalanced data does not imply a decrease in the accuracy. In these data sets, this may be explained since there is confusion in distinguishing correctly the models Erdős-Renyi (class 2) and Watts–Strogatz (class 3). Therefore, unbalancing the data in such a way that the number of units associated with these networks is lower may increase the performance of the model.

Tables 5.19, 5.20 and 5.21 display the balanced accuracy values associated with the classes 1, 2 and 3, respectively.

	OVA		OVO		CLDF w/ weights		CLDF w/o weights	
	train	test	train	test	train	test	train	test
Balanced	86.89	46.00	100.00	95.00	98.34	95.00	99.43	95.00
UnbalER	98.15	47.14	100.00	96.43	97.99	70.00	98.15	65.00
UnbalWS	97.96	61.04	97.78	95.45	94.44	100.00	99.07	95.45
UnbalBA	100.00	59.78	100.00	100.00	100.00	75.00	100.00	100.00
UnbalERWS	95.24	73.08	100.00	92.31	96.46	80.77	94.08	96.15
UnbalWSBA	83.33	32.35	100.00	100.00	100.00	97.06	88.89	100.00
UnbalERBA	99.09	47.44	100.00	100.00	94.44	100.00	94.44	100.00
UnbalAll	96.88	60.29	97.69	97.06	100.00	75.00	100.00	100.00

TABLE 5.19: Balanced accuracy for class 1 for the multi-class classification strategies (%).

Concerning class 1, both One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) capture extremely well the behaviour of this class. One-Versus-One (OVO) balanced accuracy values are consistently higher than 92%. Regarding the Consecutive Linear Discriminant Functions (CLDF) strategy according to the Definition 3.9, i.e., using weights, balanced accuracy values are consistently higher than 70%. Regarding the Consecutive Linear Discriminant Functions (CLDF) strategy according to the Definition 3.8, i.e., without using weights, balanced accuracy values are consistently higher than 65%. Even One-Versus-All (OVA), which registers the worst performance, displays balanced accuracy values above 83% in the training data set.

It is also relevant to mention that this class behaviour is perfectly captured twenty-one times.

Concerning class 2, the balanced accuracy values associated with One-Versus-All (OVA) are clearly lower than those associated with One-Versus-One (OVO) or Consecutive Linear Discriminant Functions (CLDF). In general, One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) learn this class behaviour and classify it properly in previously unseen data. Moreover, Consecutive Linear Discriminant Functions (CLDF), according to the Definition 3.8 and 3.9, tend to perform slightly worse than One-Versus-One (OVO).

When observing the balanced accuracy values associated with data set *UnbalER* for the strategy One-Versus-One (OVO), it is evident the low value in the testing data set:

					CLDF		CLDF	
	OVA		OVO		w/ weights		w/o weights	
	train	test	train	test	train	test	train	test
Balanced	50.88	40.38	97.22	95.45	89.95	82.34	92.73	85.49
UnbalER	65.93	86.96	98.88	47.83	84.22	86.96	82.48	82.61
UnbalWS	63.41	56.41	83.41	79.49	90.00	95.83	93.41	96.15
UnbalBA	56.86	41.91	94.12	97.06	89.22	94.12	87.25	100.00
UnbalERWS	63.64	90.00	99.06	93.33	89.79	83.33	86.19	96.67
UnbalWSBA	51.57	58.79	95.90	92.86	91.18	96.15	80.16	85.16
UnbalERBA	55.32	33.33	99.06	93.33	94.34	86.67	92.45	100.00
UnbalAll	62.50	20.00	97.50	100.00	85.83	95.00	88.89	95.00

TABLE 5.20: Balanced accuracy for class 2 for the multi-class classification strategies (%).

47.83%. The reason behind this may be connected to the data set constitution. Recall that the *UnbalER* data set is unbalanced, because the class Erdős-Renyi (class 2) only has 12 units.

					CLDF		CLDF	
	OVA		OVO		w/ weights		w/o weights	
	train	test	train	test	train	test	train	test
Balanced	58.68	51.45	94.32	95.83	82.95	91.30	87.50	87.32
UnbalER	90.96	59.44	97.67	96.15	77.91	100.00	82.56	96.15
UnbalWS	63.71	41.30	92.70	91.30	92.75	97.83	70.00	100.00
UnbalBA	53.28	51.67	92.68	96.67	86.59	96.67	84.15	100.00
UnbalERWS	94.44	57.14	95.00	100.00	75.00	100.00	80.00	100.00
UnbalWSBA	52.40	62.50	92.79	87.50	81.25	100.00	88.46	81.25
UnbalERBA	75.45	54.17	98.86	91.67	94.32	83.33	92.05	100.00
UnbalAll	55.29	39.77	89.18	93.75	79.22	100.00	80.00	93.75

TABLE 5.21: Balanced accuracy for class 3 for the multi-class classification strategies (%).

Regarding class 3, both One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) perform better than One-Versus-All (OVA). The behaviour of this class is perfectly captured nine times. Moreover, these perfectly balanced accuracy values tend to be associated with the Consecutive Linear Discriminant Functions (CLDF) strategies.

The balanced accuracy values associated with strategy One-Versus-One (OVO) are consistently higher than 87%, the ones associated with strategy Consecutive Linear Discriminant Functions (CLDF), according to the Definition 3.9, i.e., using weights, are consistently higher than 75% and the ones associated with strategy Consecutive Linear Discriminant Functions (CLDF), according to the Definition 3.8, i.e., without using weights, are consistently higher than 70%. One-Versus-One (OVO) tends to be more stable since the generalisation worsens no more than 8%.

Although the balanced accuracy was the measure used, we realise that, in the presence of unbalanced data sets, the balanced accuracy smooths the higher error rate that may be associated with the minority classes. For example, as referenced previously, in Table 5.4a, regarding the number of units that belong to class 2, we noticed that approximately 63.64% (7/11) were classified as belonging to the other classes and, in Table 5.20, we notice a balanced accuracy of 65.93%. Regarding class 1, 0% were misclassified and the balanced accuracy registered, in Table 5.19, is 98.15%. Finally, regarding class 3, 9.30% (4/43) were classified as belonging to class 2 and the balanced accuracy registered, in Table 5.21, is 90.96%. Although the error rate associated with class 2 is much higher than the ones associated with the remaining classes, the balanced accuracy values do not register such a large difference between the classes. Therefore, when drawing conclusions based on the balanced accuracy values we must keep in mind that the results associated with the minority classes may be mitigated.

Tables 5.22 and 5.23 display the average number of samples performed in the binary subproblems and the exact number of samples performed in the majority vote part of the classification. The number of sample operations performed suggests the level of randomness in the classification process.

	OVA		OVO		CLDF	CLDF
	binary	MV	binary	MV	w/ weights	w/o weights
Balanced	0	105	1	15	0	0
UnbalER	0	11	0	4	0	8
UnbalWS	0	56	0	28	1	2
UnbalBA	0	83	0	10	0	1
UnbalERWS	0	8	1	4	3	0
UnbalWSBA	0	51	1	4	0	4
UnbalERBA	0	17	1	4	0	0
UnbalAll	0	64	5	6	13	0

TABLE 5.22: Number of samples performed in the majority vote for the multi-class classification strategies on the train data set of network data, including the binary subproblems' average number of samples.

Both One-Versus-One (OVO) and One-Versus-All (OVA) display a low number of samples performed in the binary subproblems of the training data set. In general, the Consecutive Linear Discriminant Functions (CLDF) strategy, when considering the Definition 3.9, i.e., using weights, also displays this behaviour, but the number of sample operations performed increases when considering the data set $UnbalAll$. Therefore, for this strategy, unbalancing the data sets appears to gradually introduce randomness to the classification process in the training step. Moreover, the Consecutive Linear Discriminant Functions (CLDF) strategy, when considering the Definition 3.8, i.e., without using weights, also displays a fairly low amount of sample operations performed.

One-Versus-All (OVA) registers a very high number of samples in the majority vote, introducing a lot of randomness in the classification process.

When comparing the data sets, it seems that the Erdős-Renyi model introduces a lot of randomness. This stems from the fact that $UnbalER$ has a lower number of sample operations performed, when using the strategies One-Versus-All (OVA) and One-Versus-One (OVO), compared to $UnbalWS$ and $UnbalBA$.

	OVA		OVO		CLDF	CLDF
	binary	MV	binary	MV	w/ weights	w/o weights
Balanced	35	20	0	2	0	0
UnbalER	24	21	0	4	0	0
UnbalWS	25	14	0	13	1	0
UnbalBA	25	14	0	3	0	0
UnbalERWS	16	8	0	2	0	0
UnbalWSBA	20	9	0	2	0	6
UnbalERBA	16	11	0	2	1	0
UnbalAll	19	8	0	1	1	0

TABLE 5.23: Number of samples performed in the majority vote for the multi-class classification strategies on the test data set of network data, including the binary subproblems' average number of samples.

Although One-Versus-All (OVA) does not register any number of samples performed in the binary subproblems of the training data sets, this number registers a steep increase, when in presence of the testing data set. By analysing Table 5.23, we can say that One-Versus-All (OVA) is mainly the result of random processes. As a consequence, the accuracy and balanced accuracy values are not meaningful.

One-Versus-One (OVO) displays less randomness. Although, performing 13 sample operations in the majority vote process is fairly high, given the testing data set's dimensions. Moreover, Consecutive Linear Discriminant Functions (CLDF), when considering both definitions, also show low values of sample operations performed.

5.3 Modelling Internet Data

5.3.1 Description of the Data

This data is related to Internet traffic redirection that is the result of Border Gateway Protocol (BGP) hijacking attacks. Border Gateway Protocol (BGP) is a gateway protocol that is used so that the Internet can exchange routing information between autonomous systems. In other words, Border Gateway Protocol (BGP) enables systems to interact and communicate with each other. A sequence of systems is called a route.

This protocol does not support security mechanisms, leading the way to attacks. These attacks induce a traffic redirection that may cause flawed routing information on a worldwide scale. It was proposed a methodology for the detection of these attacks based on a set of probes that are spread worldwide [29]. They not only periodically measure the Round-Trip Time (RTT) and trace routes to targets, but also report their measurements. Round-Trip Time (RTT) is the time interval that the packages take to follow the route $probe \longrightarrow target \longrightarrow probe$.

By studying these measurements in several probes, it is possible to detect traffic redirection. Traffic can be classified as either:

- no relay or regular: this is what is expected in a normal situation.

$$probe \longrightarrow target \longrightarrow probe$$

- under attack: when under attack, the traffic is deviated before reaching the target.

$$probe \longrightarrow relay \longrightarrow target \longrightarrow probe$$

What is here on out referenced as *target* is the autonomous system that receives data and sends it to the respective *probe*. A *probe* is a machine from which data packages are sent, received and measures are assessed. These measures are useful to decide if an observation is atypical or not. A *relay* concerns the attacker that deviates the traffic from probe to target.

In the data, we considered four targets (Chicago, Frankfurt, Hong Kong and London), twelve probes and four relays. Moreover, the classification problem relies on identifying the relays (5 classes):

- 0 - no relay (regular);

- 1 - relay in Los Angeles;
- 2 - relay in Madrid;
- 3 - relay in Moscow;
- 4 - relay in São Paulo;

5.3.2 Data sets

Based on the structure proposed, the data sets used in this thesis were obtained. There are four distinct data sets with a different number of observations, but the same number of histogram-valued variables. Each histogram-valued variable is associated with information collected from a specific probe. There are 12 probes, therefore there are 12 histogram-valued variables. Every 120 seconds, each probe sends 10 packets, measures 10 Round-Trip Time (RTT), and records minimum, median and maximum. Table 5.24 displays the targets, probes and relays used in these data sets.

Targets	Probes			Relays
Chicago1	Amsterdam	Iceland	SaoPaulo2	LA1
Frankfurt1	Chicago2	Israel	Johannesburg1	Madrid
Hong Kong	VdM	LA2	Johannesburg2	Moscow
London	Frankfurt2	Milan	Sweden	SaoPaulo1

TABLE 5.24: Targets, probes and relays used in the data sets.

Each dataset is referenced in this thesis as:

1. data.T1.all: data measured, by all 12 probes, for target 1 (Chicago) ;
2. data.T2.all: data measured, by all 12 probes, for target 2 (Frankfurt);
3. data.T3.all: data measured, by all 12 probes, for target 3 (Hong Kong);
4. data.T4.all: data measured, by all 12 probes, for target 4 (London).

The description of each dataset is given in Table 5.25. Notice that the classes are unbalanced. The data associated with class 0, i.e., no relay (regular) clearly display the majority of the units in the data sets.

Each data set was fairly heavy, hence, with the function *sample* in *RStudio*, only a sample was used to run the strategies developed. The description of each sampled dataset is given in Table 5.26.

	0	1	2	3	4
data.T1.all	8732	848	598	768	914
data.T2.all	8663	745	538	832	798
data.T3.all	8544	770	564	808	810
data.T4.all	8569	681	567	782	835

TABLE 5.25: Number of units per relay in each Internet dataset.

	0	1	2	3	4
data.T1.all	271	30	23	20	16
data.T2.all	285	22	13	26	22
data.T3.all	251	21	12	21	29
data.T4.all	265	22	15	22	34

TABLE 5.26: Number of units per relay in each sampled Internet dataset.

As it is possible to observe we only worked with a small sample of the original data set. This is due to the execution time run that will be discussed in the following subsection.

5.3.3 Discussion of Results

Concerning the sampled data, we additionally divided into training and testing data sets considering 80% and 20%, respectively. Since we have 5 classes, the One-Versus-All (OVA) strategy required 5 binary subproblems, but One-Versus-One (OVO) strategy required 10 binary subproblems.

Tables 5.27, 5.28, 5.29 and 5.30 display the confusion matrices obtained from applying the strategies One-Versus-All (OVA), One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF), considering the Definitions 3.9 (with weights) and 3.8 (without weights), respectively, to the four sampled Internet data sets.

Regarding the One-Versus-All (OVA) strategy, it is clear that the behaviour of class 0 (regular) is not properly learned. Moreover, some peculiarities are clearly identified. With regards to the data set *data.T1.all*, this strategy does not properly classify any unit associated with classes 2 and 4. In the data set *data.T2.all* and *data.T4.all*, all units from class 3 can be classified by the strategy used with every class except class 2. Regarding *data.T2.all* every unit that belongs to class 4 is not classified properly. In data *data.T3.all* this behaviour is associated with classes 1 and 2. Finally, in the data *data.T4.all* this behaviour is associated with classes 1 and 4.

		Reference				
		0	1	2	3	4
Prediction	0	12	1	0	0	1
	1	10	2	2	1	2
	2	16	1	0	0	0
	3	9	1	3	1	2
	4	13	4	1	5	0

(A) data.T1.all.

		Reference				
		0	1	2	3	4
Prediction	0	16	0	0	1	1
	1	8	1	2	1	0
	2	16	2	1	0	1
	3	14	2	1	1	1
	4	10	1	1	1	0

(B) data.T2.all.

		Reference				
		0	1	2	3	4
Prediction	0	11	1	0	0	2
	1	12	0	0	0	1
	2	5	1	0	1	1
	3	9	0	2	3	0
	4	13	0	2	0	1

(C) data.T3.all.

		Reference				
		0	1	2	3	4
Prediction	0	11	1	0	2	0
	1	13	0	0	2	2
	2	11	2	1	0	3
	3	15	0	0	2	1
	4	10	0	0	2	0

(D) data.T4.all.

TABLE 5.27: Confusion matrices that result of the model applied to the Internet testing data set using OVA.

		Reference				
		0	1	2	3	4
Prediction	0	60	0	0	0	0
	1	0	9	0	0	0
	2	0	0	6	0	0
	3	0	0	0	7	0
	4	0	0	0	0	5

(A) data.T1.all.

		Reference				
		0	1	2	3	4
Prediction	0	64	0	0	0	0
	1	0	6	0	0	0
	2	0	0	5	0	0
	3	0	0	0	4	0
	4	0	0	0	0	3

(B) data.T2.all.

		Reference				
		0	1	2	3	4
Prediction	0	43	0	1	0	0
	1	0	2	2	0	0
	2	2	0	0	0	0
	3	3	0	1	4	0
	4	2	0	0	0	5

(C) data.T3.all.

		Reference				
		0	1	2	3	4
Prediction	0	59	0	0	0	0
	1	1	3	0	0	0
	2	0	0	1	0	0
	3	0	0	0	8	0
	4	0	0	0	0	6

(D) data.T4.all.

TABLE 5.28: Confusion matrices that result of the model applied to the Internet testing data set using OVO.

In respect of the One-Versus-One (OVO) strategy, the *data.T1.all* and *data.T2.all* are perfectly classified. When observing the results associated with the *data.T4.all*, we realise that only one unit, from class 0, is miss classified, as belonging to class 1.

Although the *data.T3.all* data set is still associated with a fairly high amount of miss classified units, it is less than the confusion observed in One-Versus-All (OVA) strategy.

		Reference				
		0	1	2	3	4
Prediction	0	60	0	0	0	0
	1	0	9	0	0	0
	2	0	0	6	0	0
	3	0	0	0	7	0
	4	0	0	0	0	5

(A) data.T1.all.

		Reference				
		0	1	2	3	4
Prediction	0	64	0	0	0	0
	1	0	6	0	0	0
	2	0	0	5	0	0
	3	0	0	0	4	0
	4	0	0	0	0	3

(B) data.T2.all.

		Reference				
		0	1	2	3	4
Prediction	0	7	0	0	0	2
	1	11	2	1	0	1
	2	13	0	1	1	0
	3	1	0	1	2	0
	4	18	0	1	1	2

(C) data.T3.all.

		Reference				
		0	1	2	3	4
Prediction	0	1	0	0	0	0
	1	0	3	0	0	0
	2	59	0	1	0	0
	3	0	0	0	7	0
	4	0	0	0	1	6

(D) data.T4.all.

TABLE 5.29: Confusion matrices that result of the model applied to the Internet testing data set using CLDF with weights.

By observing Table 5.29, we conclude that Consecutive Linear Discriminant Functions (CLDF) with weights strategy performs perfectly in *data.T1.all* and *data.T2.all*. Regarding *data.T4.all*, it is possible to understand that the vast majority of units from class 0 are misclassified as belonging to class 2. Once again, the confusion in *data.T3.all* is noticeable.

Finally, by observing Table 5.30, we conclude that Consecutive Linear Discriminant Functions (CLDF), according to the Definition 3.8, i.e., without using weights, performs perfectly in *data.T1.all* and *data.T2.all*. This strategy produces similar results to the ones presented in Table 5.29.

By analysing Table 5.31, we realise that both One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) perform perfectly for the data set *data.T1.all* and *data.T2.all*, showing an accuracy of 100%, without performing a single sample operation, concerning the testing data set.

When considering *data.T3.all*, the accuracy of One-Versus-One (OVO) is fairly good. Regarding the other methods, the best one uses the One-Versus-All (OVA) approach with an associated accuracy that is a little over 23%.

		Reference				
		0	1	2	3	4
Prediction	0	60	0	0	0	0
	1	0	9	0	0	0
	2	0	0	6	0	0
	3	0	0	0	7	0
	4	0	0	0	0	5

(A) data.T1.all.

		Reference				
		0	1	2	3	4
Prediction	0	64	0	0	0	0
	1	0	6	0	0	0
	2	0	0	5	0	0
	3	0	0	0	4	0
	4	0	0	0	0	3

(B) data.T2.all.

		Reference				
		0	1	2	3	4
Prediction	0	7	0	0	0	2
	1	11	2	1	0	1
	2	13	0	1	1	0
	3	1	0	1	2	0
	4	18	0	1	1	2

(C) data.T3.all.

		Reference				
		0	1	2	3	4
Prediction	0	1	0	0	0	0
	1	0	3	0	0	0
	2	59	0	1	0	0
	3	0	0	0	7	0
	4	0	0	0	1	6

(D) data.T4.all.

TABLE 5.30: Confusion matrices that result of the model applied to the Internet testing data set using CLDF without weights.

			CLDF	CLDF
	OVA	OVO	w/ weights	w/o weights
data.T1.all	17.24	100.00	100.00	100.00
data.T2.all	23.17	100.00	100.00	100.00
data.T3.all	23.08	83.08	21.54	21.54
data.T4.all	17.95	98.72	23.08	23.08

TABLE 5.31: Accuracy for the multi-class classification strategies for the Internet testing data sets (%).

Finally, *data.T4.all* is properly modelled by One-Versus-One (OVO), registering an accuracy value over 98%.

In general, we conclude that One-Versus-All (OVA) strategy does not seem to produce models that properly fit the data. The best model produced with this strategy registers an accuracy of 23.17% which is unacceptable.

In general, a lot of randomness is associated with the strategy One-Versus-All (OVA) in the testing data set. The evidence is displayed in Table 5.32 since the average number of sample operations performed in the binary subproblems is equal or very close to the dimension of the testing data set. With the exception of the strategy One-Versus-One (OVO) over the data set *data.T3.all*, the remaining strategies display a scarce number of sample operations performed.

	OVA		OVO		CLDF w/ weights	CLDF w/o weights
	binary	MV	binary	MV	MV	MV
data.T1.all	87	79	0	0	0	0
data.T2.all	82	64	0	0	0	0
data.T3.all	65	56	0	14	0	0
data.T4.all	78	65	0	1	0	0

TABLE 5.32: Number of samples performed for the multi-class classification strategies on the test data set of the Internet data set, including the binary subproblems' average number of samples.

	Class	OVA	OVO	CLDF w/ weights	CLDF w/o weights
data.T1.all	0	56.30	100.00	100.00	100.00
	1	51.50	100.00	100.00	100.00
	2	39.51	100.00	100.00	100.00
	3	47.77	100.00	100.00	100.00
	4	35.98	100.00	100.00	100.00
data.T2.all	0	56.94	100.00	100.00	100.00
	1	51.10	100.00	100.00	100.00
	2	47.66	100.00	100.00	100.00
	3	50.96	100.00	100.00	100.00
	4	41.77	100.00	100.00	100.00
data.T3.all	0	51.00	89.67	50.33	50.33
	1	39.68	98.41	89.68	89.68
	2	43.44	48.36	51.03	51.03
	3	78.48	96.72	73.36	73.36
	4	47.50	98.33	53.33	53.33
data.T4.all	0	50.83	99.17	50.83	50.83
	1	38.67	99.33	100.00	100.00
	2	89.61	100.00	61.69	61.69
	3	51.07	100.00	93.75	93.75
	4	41.67	100.00	99.31	99.31

TABLE 5.33: Balanced accuracy per class for the multi-class classification strategies (%).

Table 5.33 displays the balanced accuracy values associated with all classes in the testing data sets. In general, One-Versus-All (OVA) is unable to properly capture the behaviour associated with each class. The balanced accuracy values are clearly lower when compared to the ones concerning One-Versus-One (OVO).

With the exception of *data.T3.all*, One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) seem to adjust fairly well to every class behaviour in the Internet data sets.

By observing Table 5.34, it is possible to understand that, even for a small portion of

	OVA	OVO	CLDF w/ weights	CLDF w/o weights
data.T1.all	2.35	1.94	0.25	0.25
data.T2.all	2.27	2.16	0.38	0.35
data.T3.all	2.00	2.03	0.43	0.40
data.T4.all	2.14	2.41	0.73	0.68

TABLE 5.34: Execution time, in hours, associated with each sampled Internet data set.

the original data sets, the execution time is fairly high. The best time runs are associated with the Consecutive Linear Discriminant Functions (CLDF) strategies. These consistently take less than an hour. Both One-Versus-One (OVO) and One-Versus-All (OVA) display execution time runs that are longer than two times the time it takes to run the Consecutive Linear Discriminant Functions (CLDF) approaches.

Moreover, there were attempts to assess the performance of the strategies in a larger sample of the data, particularly for 10% of the data. However, the execution time run was immense. For a sample of 10% of the data associated with $data.T1.all$, regarding the strategy One-Versus-All (OVA), the execution time run is higher than 35.70 hours. At this time, the strategy One-Versus-All (OVA) was running the first binary subproblem. Regarding the strategy One-Versus-One (OVO), the execution time run is higher than 23.41 hours. At this time, the strategy One-Versus-One (OVO) was also running the first binary subproblem. Regarding the strategy Consecutive Linear Discriminant Functions (CLDF), concerning both the Definitions 3.8 (without weights) and 3.9 (with weights) after 11.63 hours the execution was completed, registering a perfect accuracy without performing sample operations.

5.4 Modelling Car Data

5.4.1 Description of the Data

This data was examined in order to display an application with interval-valued variables. This data consists of characteristics of car models. The classification problem relies on identifying the model class (4 classes):

- 1 - Utilitarian;
- 2 - Berlina;

- 3 - Luxury;
- 4 - Sportive;

The description of the classes in this dataset is given in Table 5.35.

Utilitarian	Berlina	Luxury	Sportive
10	8	8	7

TABLE 5.35: Number of units per class in the cars dataset.

5.4.2 Data set

This data set is fairly small. It has 33 units and 4 interval-valued variables:

- Price: it indicates how expensive a car is;
- Engine Capacity: it indicates how powerful the car is;
- Top Speed: it indicates the fastest the car can go;
- Acceleration: it indicates the rate that a car increases its speed;

5.4.3 Discussion of Results

Since this data set is very small, we performed leave-one-out cross-validation. In each case, in the testing unit, we evaluate if the class is correctly predicted.

		CLDF	CLDF
OVA	OVO	w/ weights	w/o weights
51.52	96.97	-	-

TABLE 5.36: Mean accuracy for the multi-class classification strategies for the cars data set (%).

Table 5.36 displays the mean accuracy of each strategy. We conclude that One-Versus-All (OVA) is the strategy that displays the worst performance since it produces models with an average accuracy of 51.52%. One-Versus-One (OVO) displays the best strategy, with a mean accuracy of 96.97%.

The results associated with the Consecutive Linear Discriminant Functions (CLDF) strategies are not displayed since the optimisation process declares either primal infeasibility or is stuck at an edge of dual feasibility, giving up.

Validation	OVA		OVO		CLDF	CLDF
set	binary	MV	binary	MV	w/ weights	w/o weights
1	0	1	0	1	-	-
2	0	0	0	1	-	-
3	0	0	0	0	-	-
4	0	1	0	1	-	-
5	0	1	0	0	-	-
6	0	0	0	0	-	-
7	0	0	0	0	-	-
8	0	1	0	0	-	-
9	0	0	0	0	-	-
10	0	1	0	0	-	-
11	0	1	0	0	-	-
12	0	1	0	0	-	-
13	0	1	0	0	-	-
14	0	1	0	0	-	-
15	0	1	0	0	-	-
16	0	1	0	0	-	-
17	0	1	0	0	-	-
18	0	1	0	0	-	-
19	0	0	0	0	-	-
20	0	1	0	0	-	-
21	0	1	0	0	-	-
22	0	1	0	0	-	-
23	0	1	0	0	-	-
24	0	1	0	0	-	-
25	0	1	0	0	-	-
26	0	0	0	0	-	-
27	0	1	0	0	-	-
28	0	1	0	0	-	-
29	0	1	0	0	-	-
30	0	1	0	0	-	-
31	0	1	0	0	-	-
32	0	0	0	1	-	-
33	0	0	0	1	-	-

TABLE 5.37: Number of samples performed in the majority vote for the multi-class classification strategies on the cars' data, including the binary subproblems' average number of samples.

Table 5.37 displays the average number of samples performed in the binary subproblems and the exact number of samples performed in the majority vote part of the classification in the validation set. Since we performed leave-one-out, the validation set is composed exactly of one unit.

In general, One-Versus-All (OVA) performs more sample operations and the unit's classification tends to be subjected to one sample operation usually in the majority vote. One-Versus-One (OVO) performs a scarce number of sample operations and, when it is performed, it tends to be in the majority vote.

5.5 General Observations

In summary, we can conclude that Consecutive Linear Discriminant Functions (CLDF) is the most deterministic strategy, i.e., it is the method that performs the lowest number of *sample* operations. Furthermore, One-Versus-One (OVO) does not tend to perform a significant number of samples. One-Versus-All (OVA) is, by far, the method that executes the most randomised operations. It is fair to assume that the number of samples performed is correlated to the performance of the model. A higher number of sample operations tends to be associated with models with lower accuracy. Moreover, it stands to reason that data that does not display clear and distinct classes are inclined to execute more sample operations.

Although unbalancing data sets does not imply a decrease in accuracy, the error rates associated with the minority classes tend to be higher than the ones associated with the remaining classes.

Some conclusions can also be drawn about the performance of the three strategies developed for multi-class classification. One-Versus-All (OVA) displays consistently the worst results, occasionally providing inaccurate classifications. In general, both One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) show stellar results.

Chapter 6

Conclusion

There has been an increase in the development of new approaches to cope with the higher power of data storage. One of these approaches is Symbolic Data Analysis (SDA) which comes with several advantages and disadvantages. Among the disadvantages, it is important to point out that most existing concepts and models developed for classic data become no longer viable. Nevertheless, one advantage stands out. It concerns the possibility of retaining the inherent variability associated with the data. We have, therefore, more informative and valuable results.

This thesis explored the possibility of developing a model to properly assess the class to which a unit belongs to in a multi-class classification problem. It focuses on linear discriminant analysis, in a multi-class setting, extending the binary classification method developed in [1].

6.1 Conclusion and Perspectives

The proposed extension of the discriminant method for histogram-valued variables allows for addressing multi-class classification problems associated with symbolic data.

On the one hand, we can reduce the multi-class classification problem to several binary subproblems. Classification is performed in two steps. Firstly, we use the Mallows distance between the scores of each unit and the barycentric scores of each class in the binary subproblems. Then, this is used to build the multi-class classification by computing the majority vote. The approaches developed in this thesis that lie on this idea are One-Versus-One (OVO) and One-Versus-All (OVA).

On the other hand, the problem may be addressed by constructing several linear discriminant functions (Consecutive Linear Discriminant Functions (CLDF)). The optimisation requires additional conditions in order to guarantee that the successive discriminant functions are uncorrelated. The classification is based on the Mallows distance between the scores of each unit and the barycentric scores of each class.

The performance of these three strategies is comparatively consistent. Even though One-Versus-All (OVA) shows the worst results, One-Versus-One (OVO) and Consecutive Linear Discriminant Functions (CLDF) tend to perform very well. However, every strategy is inadequate when the data for which we aim to provide a multi-class classification does not display a clusterable structure.

To conclude, the three techniques yield a way of performing multi-class classification of symbolic data that has not been executed until now. These procedures are likely to be useful in classification problems, with more than two *a priori* classes, where the variability associated with the data is crucial.

6.2 Limitations and Future Work

Currently, the main issue identified was the heavy computations required. For large data sets, either with a large amount of histogram-valued variables or units, the time required to run the linear discriminant analysis is immense, especially for the strategies One-Versus-All (OVA) and One-Versus-One (OVO). Moreover, the time it takes to run the analysis is also sensitive to the number of subintervals of the realisations of the histogram-valued variables. For this reason, improving the code time run is in great demand.

Furthermore, the software used to perform the optimisation step may not be the most appropriate. It would be of interest to study options that make use of software tools that were developed entirely for mathematical optimisation.

Bibliography

[1] S. Dias, P. Brito, and P. Amaral, "Discriminant analysis of distributional data via fractional programming", *European Journal of Operational Research*, vol. 294, no. 1, pp. 206–218, 2021. DOI: https://doi.org/10.1016/j.ejor.2021.01.025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221721000424.

[2] J. Myers. "This is how much data we're using on our phones". (Aug. 2021), [Online]. Available: https://www.weforum.org/agenda/2021/08/how-the-pandemic-sparked-a-data-boom/ (visited on 04/17/2023).

[3] A. Irpino and R. Verde, "Linear regression for numeric symbolic variables: A least squares approach based on Wasserstein distance", *Advances in Data Analysis and Classification*, vol. 9, pp. 81–106, 2015. DOI: https://doi.org/10.1007/s11634-015-0197-7.

[4] S. Dias and P. Brito, "Linear regression model with histogram-valued variables", *Statistical Analysis and Data Mining*, vol. 8, no. 2, pp. 75–113, 2015. DOI: https://doi.org/10.1002/sam.11260.

[5] P. Rahman, B. Beranger, S. Sisson, and M. Roughan, "Likelihood-based inference for modelling packet transit from thinned flow summaries", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 571–583, 2022. DOI: https://doi.org/10.1109/TSIPN.2022.3188457.

[6] S. Dias, "Linear Regression with Empirical Distributions", Ph.D. dissertation, Universidade do Porto, 2014.

[7] A. Irpino and R. Verde, "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data", in *Data Science and Classification*, V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, Eds., Berlin, Heidelberg: Springer Berlin

Heidelberg, 2006, pp. 185–192. DOI: https://doi.org/10.1007/3-540-34416-0_20.

[8] A. Irpino and R. Verde, “Basic statistics for distributional symbolic variables: A new metric-based approach”, *Advances in Data Analysis and Classification*, vol. 9, pp. 143–175, 2015. DOI: https://doi.org/10.1007/s11634-014-0176-4.

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

[10] P. Brito, “Symbolic data analysis: Another look at the interaction of data mining and statistics”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 281–295, 2014. DOI: https://doi.org/10.1002/widm.1133.

[11] L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, 2006.

[12] J. Arroyo, “Métodos de Predicción para Series Temporales de Intervalos e Histogramas”, Ph.D. dissertation, Universidad Pontificia Comillas, Madrid, Espanha, 2008.

[13] G. Casella and R. Berger, *Statistical Inference* (Duxbury advanced series). Brooks/-Cole Publishing Company, 1990. [Online]. Available: https://books.google.pt/books?id=nA%5C_vAAAAMAAJ.

[14] P. Prakash and M. R. Sertel, “Semilinear (topological) spaces and applications”, Massachusetts Inst. of Tech. Cambridge, MA, United States, Tech. Rep., 1971.

[15] C. Mallows, “A note on asymptotic joint normality”, *The Annals of Mathematical Statistics*, vol. 43, no. 2, pp. 508–515, 1972.

[16] J. Arroyo and C. Maté, “Forecasting histogram time series with k-nearest neighbours methods”, *International Journal of Forecasting*, vol. 25, no. 1, pp. 192–207, 2009. DOI: https://doi.org/10.1016/j.ijforecast.2008.07.003.

[17] L. Billard and E. Diday, “From the statistics of data to the statistics of knowledge: Symbolic data analysis”, *Journal of the American Statistical Association*, vol. 98, pp. 470–487, Jun. 2003. DOI: https://doi.org/10.1198/016214503000242.

[18] P. Amaral, I. Bomze, and J. Júdice, “Copositivity and constrained fractional quadratic problems”, *Mathematical Programming*, vol. 146, pp. 325–350, Aug. 2014. DOI: https://doi.org/10.1007/s10107-013-0690-8.

[19] L. Vandenberghe and S. Boyd, "Semidefinite programming", *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996. DOI: https://doi.org/10.1137/1038003.

[20] A. Band. "Multi-class classification — one-vs-all & one-vs-one". (May 2020), [Online]. Available: https://towardsdatascience.com/multi-class-classification-one-vs-all-one-vs-one-94daed32a87b.

[21] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes", *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011. DOI: https://doi.org/10.1016/j.patcog.2011.01.017.

[22] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, "Analyzing the presence of noise in multi-class problems: Alleviating its influence with the one-vs-one decomposition", *Knowledge and Information Systems*, vol. 38, pp. 179–206, 2014. DOI: https://doi.org/10.1007/s10115-012-0570-1.

[23] D. Morrison, *Multivariate Statistical Methods* (McGraw-Hill engineering reference guide series). McGraw-Hill, 1967. [Online]. Available: https://books.google.pt/books?id=kAJRAAAAMAAJ.

[24] S. Sharma, *Applied Multivariate Techniques*. Wiley, 1996.

[25] A. Irpino, *Histdawass: Histogram-valued data analysis*, R package version 1.0.6, 2021. [Online]. Available: https://CRAN.R-project.org/package=HistDAWass.

[26] H. Corrada Bravo and B. Borchers, *Rcsdp: R interface to the csdp semidefinite programming library*, R package version 0.1.57.2, 2021. [Online]. Available: https://CRAN.R-project.org/package=Rcsdp.

[27] F. Schwendinger, *Roi.plugin.alabama: 'alabama' plug-in for the R optimization infrastructure*, R package version 1.0-0, 2020. [Online]. Available: https://CRAN.R-project.org/package=ROI.plugin.alabama.

[28] A. Colombo and R. Jaarsma, "A powerful numerical method to combine random variables", *IEEE Transactions on Reliability*, vol. R-29, no. 2, pp. 126–129, 1980. DOI: https://doi.org/10.1109/TR.1980.5220750.

[29] P. Salvador and A. Nogueira, "Customer-side detection of internet-scale traffic redirection", in *2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, IEEE, 2014, pp. 1–5. DOI: https://doi.org/10.1109/NETWKS.2014.6958532.

[30] M. Posfai and A.-L. Barabasi, *Network Science*. Citeseer, 2016.

[31] S. Jalan and C. Sarkar, “Complex networks: An emerging branch of science”, *Phys. News*, vol. 47, pp. 3–4, 2017.

[32] H. Lavicka, “Simulations of Agents on Social Network”, Ph.D. dissertation, Czech Technical University in Prague, 2010.