Advances in Data Analysis and Classification (2025) 19:1119–1146
https://doi.org/10.1007/s11634-025-00624-x

REGULAR ARTICLE

# Parametric models for distributional data

**Paula Brito**[1,2] · **A. Pedro Duarte Silva**[3]

Received: 6 March 2024 / Revised: 10 December 2024 / Accepted: 14 January 2025 /
Published online: 10 March 2025


**Abstract**
We present parametric probabilistic models for numerical distributional variables. The proposed models are based on the representation of each distribution by a location measure and inter-quantile ranges, for given quantiles, thereby characterizing the underlying empirical distributions in a flexible way. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance–covariance matrix. For all cases, maximum likelihood estimators of the corresponding parameters are derived. This modelling allows for hypothesis testing and multivariate parametric analysis. The proposed framework is applied to Analysis of Variance and parametric Discriminant Analysis of distributional data. A simulation study examines the performance of the proposed models in classification problems under different data conditions. Applications to Internet traffic data and Portuguese official data illustrate the relevance of the proposed approach.

**Keywords** Analysis of variance · Discriminant analysis · Histogram data · Symbolic data

**Mathematics Subject Classification** 62R07 · 62H30

## 1 Introduction

In classical multivariate statistics, data mining, and machine learning, data is represented in an array where each row represents an unit, for which one single value is

✉ Paula Brito
mpbrito@fep.up.pt

A. Pedro Duarte Silva
psilva@ucp.pt

1 Faculdade de Economia, Universidade do Porto, Porto, Portugal

2 LIAAD, INESC TEC, Porto, Portugal

3 Católica Porto Business School and CEGE, Universidade Católica Portuguesa, Porto, Portugal

recorded for each numerical or categorical variable (in columns). This representation model is however limited when the data to be analysed comprises variability. That is the case when the units under analysis are not single elements, but general concepts, such as car models, rather than specific vehicles, or groups formed on the basis of some given common properties. This is very pertinent in data mining applications where huge sets of data are collected, but data should be analysed at a higher level—e.g. take the case of large department stores, which record data on each purchase made (amount spent, items purchased, etc.), but where the focus does not lay on individual purchases but rather on consumer behaviour, and therefore information about the purchases of each client, or specific groups of clients, must be somehow aggregated. Then, for each descriptive variable, the observed variability inherent to each concept or group should be taken into account, not just relying, as it is traditionally done, on central tendency measures (means, medians, modes), to avoid an important loss of pertinent information. In the era of Big Data, there is increasing demand for methodologies capable of reducing the original dataset to a manageable level, while keeping all relevant information. Symbolic Data Analysis (see e.g. Bock and Diday 2000; Billard and Diday 2003, 2006; Brito 2014) provides a framework for the representation and analysis of such data. This approach allows for data aggregation at different degrees of granularity, without ignoring the corresponding variability. New variable types have been introduced whose realizations are not single real values or categories, but rather finite sets, intervals, or, more generally, distributions over a given domain. Methods for the (multivariate) analysis of such symbolic data have been developed which allow taking into account the variability expressed in the data representation.

In this work, we focus on numerical data described by empirical distributions. These may take the form of histograms or quantile vectors. Histogram-valued data analysis has been addressed by several authors, following non-parametric approaches. In particular, relevant contributions include Dias and Brito (2015), Irpino and Verde (2015), Zhao et al. (2022) (regression); Dias et al. (2021) (discriminant analysis); Irpino and Verde (2006), Verde and Irpino (2007), Kim and Billard (2011), Brito and Chavent (2012), Irpino et al. (2014), Irpino et al. (2017), Billard and Kim (2017), De Carvalho et al. (2021) (clustering); Verde et al. (2015), Billard and Kim (2017) (dimension reduction); Arroyo and Maté (2009), Arroyo et al. (2011), Gonzalez-Rivera and Arroyo (2012) (time series and forecasting). Analysis of distributional data in the form of quantile vectors has been developed in Ichino (2011), Ichino et al. (2021), Umbleja et al. (2021), Ichino (2022). Representation and methods for the analysis of distributional data may be found in Brito and Dias (2022). Following a different approach, Jin and Billard (2022) model histogram-valued data using copulas, allowing for maximum likelihood estimation and inference.

Our aim is to propose a model allowing for the representation of empirical distributions in a flexible way, and opening the path to parametric inference methodologies. For this purpose, we need probabilistic models for numerical distributional variables. We propose parametric models based on the representation of each distribution by a central statistic $C$, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles $\psi_1, \ldots, \psi_q$. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance–covariance matrix. This assumption is rooted in classical multivariate data analysis, where

often a multivariate Normal distribution is assumed for the descriptive variables, leading to good results, even when multinormality does not strictly hold. Our proposal is inspired in the model for interval-valued data presented in Brito and Duarte Silva (2012), where parametric inference methodologies based on probabilistic models for interval variables are developed. By using a representation of empirical distributions with finer granularity, based on a central location measure and inter-quantile ranges for each descriptive distributional variable, we use more information, and may take into account the empirical distribution of the microdata in further detail. However, this may potentially lead to high dimensional models, with a large number of parameters to be estimated. By considering restricted sparser covariance configurations, where some of these parameters are set to zero, the number of estimated parameters is reduced. This allows for a more efficient estimation, and avoids overfitting, which is particularly important in the case of datasets with a small/medium number of observations.

The proposed model then allows for multivariate parametric analysis of distributional data. This clearly extends the scope of the analysis of histogram-valued data, for which, to the best of our knowledge, only non-parametric methodologies have been developed so far. In this paper, we address Analysis of Variance and Discriminant Analysis. We note that both these methods address related problems, where the population and the given sample are divided into well defined (a priori) groups. Analysis of Variance tests whether the groups have identical means, whereas Discriminant Analysis, assuming the groups are indeed different, develops classification rules for future observations of unknown origin. The model and the methods proposed in this paper are being implemented in a public-domain R package.

Researchers in the area of Functional Data Analysis (FDA) have also addressed the analysis of data as distributions. In Delicado (2011), the author addresses dimensionality reduction when data consist of density functions of one given descriptive variable, with application to the analysis of population pyramids. The authors in Petersen and Müller (2016) propose a transformation to map probability densities into a linear function space, by using a suitably chosen continuous and invertible map. The goal is to use methods for Hilbert space valued data, not directly applicable to densities. In the transformed space, FDA methodology, such as functional regression or classification, may be then implemented. In Grygar et al. (2024), the authors propose a methodology for geochemical mapping data, with post-stratification of a large dataset, where trivariate PDFs are analysed. This is based on Functional Data Analysis of probability densities in the framework of Bayes spaces. Panaretos and Zemel (2019) discuss considering the space of probability measures equipped with a Wasserstein distance for FDA. Following a Symbolic Data Analysis (SDA) point of view, our approach, is however essentially different. Our aim is to develop multivariate analysis methods, based on necessary and sufficient indicators that characterize the underlying empirical distributions in a flexible way. It is not our goal to full adjust the given (micro)data and study their behaviour as functions. Furthermore, most FDA applications concern only one, or quite few functional variables, which are observed at a (large) number of discrete points. The SDA approach typically addresses multivariate problems with a larger number of symbolic variables, which may be observed at a large or small number of points and which are somehow aggregated to obtain a description of the higher order units under analysis.

The remainder of the paper is organized as follows. Section 2 introduces the representation for distribution-valued variables. Then, in Sect. 3, we present the proposed model. Analysis of Variance is developed in Sect. 4, and Discriminant Analysis in Sect. 5. Section 6 presents a simulation study addressing Linear and Quadratic Discriminant Analysis. Section 7 describes two applications, respectively to Internet traffic data and to Portuguese official data. Section 8 concludes the paper and opens avenues for further developments.

## 2 Distributional variables

Let $S = \{s_1, \ldots, s_n\}$ be the set of $n$ units under analysis. We consider that for each unit, the descriptive variables are (in general) not constant, but present some variability, and we assume that a set of quantiles, a probability distribution, or a sample, from which quantiles may be derived, are given.

We represent the "values" of a numerical distributional variable by an ordered vector of quantiles. Formally, a numerical distributional variable is defined by an application

$$\begin{aligned} Y &: S \to T \\ s_i &\to Y(s_i) = (\psi_{1i}, \ldots, \psi_{qi}) \end{aligned}$$

where $q$ is a positive integer, $\psi_{1i} \leq \ldots \leq \psi_{qi}$ are order statistics of the relevant sample or distribution, and $T$ is the set of ordered $q$-dimensional vectors of $O \subseteq I\!R$. Here $\psi_{1i}$ and $\psi_{qi}$ are typically either the minimum and maximum, respectively, or small and large quantiles suitably chosen in order to disregard severe outliers.

Let $Q$ be an $n \times p$ matrix containing the values of $p$ numerical distributional variables on $S$. Each $s_i \in S$ is hence represented by a $p$-dimensional vector of ordered vectors, $Q_i = (Q_{i1}, \ldots, Q_{ip})^t, i = 1, \ldots, n$, with $Q_{ij} = Y_j(s_i) = (\psi_{1ij}, \ldots, \psi_{q_j ij})^t, j = 1, \ldots, p$ (see Table 1).

This representation is related to histogram-valued variables $Y$ as defined in Brito and Dias (2022). The value of such a variable is written as $Y(s_i) = \left\{I_{i1}, p_{i1}; I_{i2}, p_{i2}; \ldots; I_{im_i}, p_{im_i}\right\}$ where $I_{i\ell}$ represents the subinterval $\ell$ for the unit $s_i$, $p_{i\ell}$ is the weight associated with the subinterval $I_{i\ell}$, and $\sum_{\ell=1}^{m_i} p_{i\ell} = 1$

**Table 1** Matrix $Q$ of numerical distributional data

| | $Y_1$ | $\ldots$ | $Y_j$ | $\ldots$ | $Y_p$ |
|---|---|---|---|---|---|
| $s_1$ | $(\psi_{111}, \ldots, \psi_{q_1 11})$ | $\ldots$ | $(\psi_{11j}, \ldots, \psi_{q_j 1j})$ | $\ldots$ | $(\psi_{11p}, \ldots, \psi_{q_p 1p})$ |
| $\ldots$ | $\ldots$ | | $\ldots$ | | $\ldots$ |
| $s_i$ | $(\psi_{1i1}, \ldots, \psi_{q_1 i1})$ | $\ldots$ | $(\psi_{1ij}, \ldots, \psi_{q_j ij})$ | $\ldots$ | $(\psi_{1ip}, \ldots, \psi_{q_p ip})$ |
| $\ldots$ | $\ldots$ | | $\ldots$ | | $\ldots$ |
| $s_n$ | $(\psi_{1n1}, \ldots, \psi_{q_1 n1})$ | $\ldots$ | $(\psi_{1nj}, \ldots, \psi_{q_j nj})$ | $\ldots$ | $(\psi_{1np}, \ldots, \psi_{q_p np})$ |

with $m_i$ the number of subintervals for the $i$th unit. Our representation corresponds to histogram-valued variables where the weights $p_{i\ell}$ are fixed for all $i = 1, \ldots, n$, and the subintervals $I_{i\ell}$ are random.

## 3 Models for distributional data

Let $Y_1, \ldots, Y_p$ be the $p$ numerical distributional variables, defined on a set of units $S = \{s_1, \ldots, s_n\}$. Here we assume that all variables are represented by the same set of $q$ quantiles, and that $\psi_{1ij} < \cdots < \psi_{qij}$ (strict inequalities), $1 \le i \le n,\ 1 \le j \le p$.

The model consists in representing $Y_j(s_i)$ by

- a central statistic $C_{ij}$ (henceforth referred to as "centre"), typically the Median $Med_{ij}$ or the MidPoint $M_{ij} = \dfrac{\psi_{1ij} + \psi_{qij}}{2}$
- the $[\psi_1, \psi_2[$ range: $R_{1ij} = \psi_{2ij} - \psi_{1ij}$
- ...
- the $[\psi_{q-1}, \psi_q]$ range: $R_{mij} = \psi_{qij} - \psi_{q-1,ij}$

Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles:

- Min–Med–Max
  - $q = 3$ quantiles; $m = 2$ intervals
  - Central statistic: Median $C_{ij} = Med_{ij}$
  - $\psi_1 = Min_{ij}, \psi_2 = Med_{ij}, \psi_3 = Max_{ij}$
- Quartiles: Min-$Q_1$-Med-$Q_3$-Max
  - $q = 5$ quantiles; $m = 4$ intervals
  - Central statistic: Median $C_{ij} = Med_{ij}$
  - $\psi_1 = Min_{ij}, \psi_2 = Q_{1ij}, \psi_3 = Med_{ij}, \psi_4 = Q_{3ij}, \psi_5 = Max_{ij}$
- Five intervals, with equally-spaced quantiles
  - $q = 6$ quantiles; $m = 5$ intervals
  - Central statistic: MidPoint $C_{ij} = M_{ij}$
  - $\psi_1 = Min_{ij}, \psi_2 = F_{ij}^{-1}(0.2), \psi_3 = F_{ij}^{-1}(0.4), \psi_4 = F_{ij}^{-1}(0.6), \psi_5 = F_{ij}^{-1}(0.8), \psi_6 = Max_{ij}$

The proposed model consists in assuming that the joint distribution of the central statistic $C$ and the logarithms of the ranges $R_\ell^* = \ln(R_\ell)$, $\ell = 1, \ldots, m$, is multivariate Normal:

$$(C, R_1^*, \ldots, R_m^*) \sim N_{(m+1)p}(\mu, \Sigma)$$

$$\mu = \left[\mu_C^t, \mu_{R_1^*}^t, \ldots, \mu_{R_m^*}^t\right]^t; \Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR_1^*} & \ldots & \Sigma_{CR_m^*} \\ \Sigma_{R_1^*C} & \Sigma_{R_1^*R_1^*} & \ldots & \Sigma_{R_1^*R_m^*} \\ \ldots & \ldots & \ldots & \ldots \\ \Sigma_{R_m^*C} & \Sigma_{R_m^*R_1^*} & \ldots & \Sigma_{R_m^*R_m^*} \end{pmatrix}$$

Multivariate normality is a common assumption in classical multivariate data analysis, allowing for straightforward representation of the dependence structure, which is not as easily done with alternative distributions. We note that the multivariate Normal distribution is a theoretically model that, like all models, rarely, if ever, holds in practice. In fact, it is well known that "... in real life there is no such thing as a true model" (McLachlan 1992, p. 152). Nevertheless, there is a considerable body of literature that demonstrates that this theoretical model can be very useful under mild deviations from its assumptions, and identifies more serious deviations under which normality based techniques are no longer recommendable. Furthermore, the distributional results used in the ANOVA/MANOVA proposals made in Sect. 4 do not rely on this assumption. In the discussion of the classification rules proposed in Sect. 5 we will review the relevant literature on the effects of normality deviations, and in the examples of Sect. 7, these deviations will be assessed and confronted with the lessons from the literature. Extensions of this approach based on more flexible models, such as the Skew-Normal distribution, are currently under development.

In the most general formulation (configuration 1) we allow for non-zero correlations among all centres and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables $Y_j$ are non-correlated, but for each variable, the centre and all its log-ranges may be correlated among themselves (configuration 2); centres (respectively, log-ranges) of different variables may be correlated, but no correlation between centres and log-ranges is allowed (configuration 3); centres (respectively, each log-range) of different variables may be correlated, but no correlation between centres and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all centres and log-ranges are non-correlated (configuration 5)—see Table 2.

Configurations 2 and 3 are particular cases of configuration 1, configuration 4 is a particular case of 3, and configuration 5 is a particular case of all the others. These are illustrated in Figs. 1 and 2.

We note that in configurations 2, 3, 4 and 5, $\Sigma$ can be written as a block diagonal matrix, after possible rearrangement of rows and columns: in configuration 2 there are $p$ blocks, all $(m+1) \times (m+1)$; in configuration 3 there are two blocks, one is $p \times p$,

**Table 2** Covariance matrix configurations

| Config | Characterization | $\Sigma$ |
|---|---|---|
| 1 | Non-restricted | Non-restricted |
| 2 | $Y_j$'s non correlated | $\Sigma_{CC}$, $\Sigma_{CR^*_\ell}$, $\Sigma_{R^*_\ell C}$ and $\Sigma_{R^*_{\ell 1}R^*_{\ell 2}}$ all diagonal |
| 3 | $C$'s non-correlated with $R^*_\ell$'s | $\Sigma_{CR^*_\ell} = \Sigma_{R^*_\ell C} = 0, \ell = 1, \ldots, m$ |
| 4 | $C$'s non-correlated with $R^*_\ell, \ell = 1, \ldots, m$<br>$R^*_{\ell 1}$, $R^*_{\ell 2}$ non-correlated for $\ell 1 \neq \ell 2$ | $\Sigma_{CR^*_\ell} = \Sigma_{R^*_\ell C} = 0$<br>$\Sigma_{R^*_{\ell 1}R^*_{\ell 2}} = 0, \ell 1 \neq \ell 2$ |
| 5 | All $C$'s and $R^*_\ell, \ell = 1, \ldots, m$ are non-correlated | $\Sigma$ diagonal |

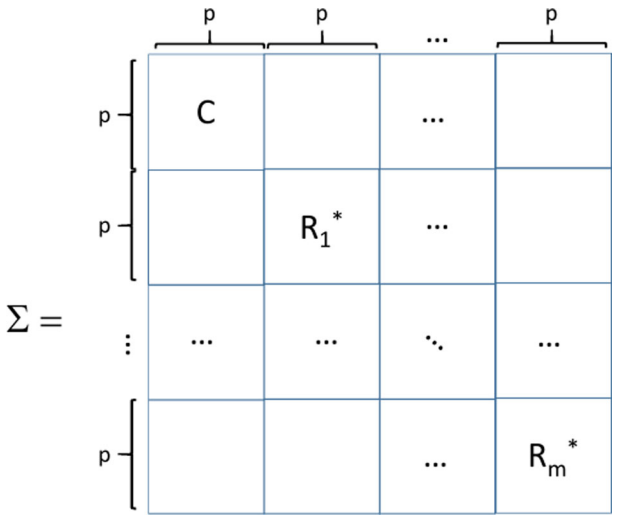


**Fig. 1** Structure of the covariance matrix

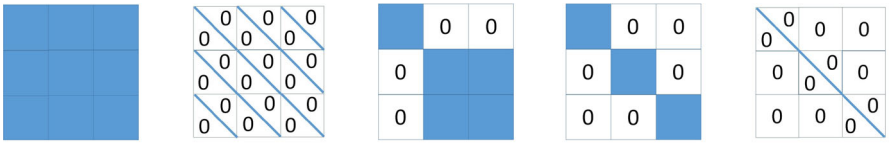


**Fig. 2** Alternative configurations for the covariance matrix for $m = 2$

and the other is $mp \times mp$; in configuration 4 there are $m + 1$ blocks, all $p \times p$, and in configuration 5 the $(m + 1)p$ blocks are single real elements.

The Gaussian model described in Duarte Silva et al. (2021) for interval-valued variables, originally proposed in Brito and Duarte Silva (2012), is a particular case of the modelling presented here, where the only quantiles considered are the minimum and the maximum, and configurations 3 and 4 collapse into a single covariance configuration.

### 3.1 Maximum likelihood estimation

Let $X_i = \left[C_i^t, R_{1i}^{*\,t}, \ldots, R_{mi}^{*\,t}\right]^t$ be the $(m+1)p$ dimensional column vector comprising all central statistics and log-ranges for unit $s_i$. Let $\bar{X}$ be sample mean of the $X_i$'s, $i = 1, \ldots, n$.

For all configurations, the log-likelihood is written as

$$ln\ L(\mu, \Sigma) = -np\ ln(2\pi) - \frac{n}{2}\ ln|\Sigma| - \frac{1}{2}\ trE\Sigma^{-1} - \frac{n}{2}\left(\bar{X} - \mu\right)^t \Sigma^{-1}\left(\bar{X} - \mu\right) \tag{1}$$

where $E = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t$.

The maximum likelihood estimators of $\mu$ and $\Sigma$ under configuration 1 are obviously the classical ones, $\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t = \frac{1}{n}E$.

For all configurations, since $\Sigma^{-1}$ is symmetric positive definite it follows that the maximum-likelihood estimate of the mean vector $\mu$ is always $\bar{X}$.

Maximization of the likelihood function with respect to $\Sigma$ then reduces to maximizing

$$ln\ L(\mu, \Sigma) = \text{ constant } - \frac{n}{2}\ ln|\Sigma| - \frac{1}{2}\ trE\Sigma^{-1} \tag{2}$$

In configurations 2, 3, 4 and 5, $\Sigma$ is subject to constraints. In these cases $\Sigma$ can be written as a block diagonal matrix, after a possible rearrangement of rows and columns. The maximum can then be obtained by separately maximizing with respect to each block $\Sigma_h$ of $\Sigma$ (Brito and Duarte Silva 2012). It is well known (see for instance Seber 1984) that (2) is maximized when $\Sigma_h = \hat{\Sigma}_h = E_h/n$, where $E_h$ is the corresponding block of matrix $E$. The maximum likelihood estimator of $\Sigma$ is then obtained with the block-wise estimators, and simply replacing by zeros the null parameters in the corresponding configuration.

In our approach, the covariance configuration is chosen by minimising the Bayesian Information Criterion (BIC) (Schwarz 1978).

## 4 Analysis of Variance

In this section we address ANOVA and MANOVA using the model introduced above.

We note that since each distributional variable $Y_j$ is modelled by the vector of $m + 1$ indicators $(C_j, R_1^*, \ldots, R_m^*)$, it follows that an Analysis of Variance of $Y_j$ is now accomplished by a $(m + 1)$-dimensional MANOVA.

Let us assume a one-way design, where the factor has $K$ levels, representing $K$ groups, and let $n_k$ be the number of observations in group $k, k = 1, \ldots, K$. Let $X_{ij} = \left[C_{ij}, R_{1ij}^*, \ldots, R_{mij}^*\right]^t$ be the $(m + 1)$ dimensional column vector comprising the central statistic and the log-ranges of variable $Y_j$ for unit $s_i$. Moreover, let $\bar{X}_{\bullet jk}$ and $\mu_{\bullet jk}$ be sample and population means of the $X_{ij}$'s in group $k$, and $\bar{X}_{\bullet j\bullet}$ the corresponding global sample mean. The null hypothesis in this case consists in stating that all $\mu_{\bullet jk}$ are equal across groups.

We follow a likelihood ratio approach as different likelihood ratio statistics $\lambda$ may be derived for each different configuration in Table 2.

The likelihood ratio statistic is $\lambda = \left( \frac{|E_{alt}|}{|E_{null}|} \right)^{\frac{n}{2}}$, where $E_{null}$ and $E_{alt}$ are $(m+1) \times (m+1)$ matrices corresponding to the null and alternative hypothesis respectively. Under the unrestricted configuration 1, $E_{null}$ and $E_{alt}$ are obviously equal to the classical ones, i.e., the sums of squares and cross-products MANOVA matrices. For configurations 2, 3, 4, and 5, $E_{null}$ is obtained from $E_j = \sum_{i=1}^{n} (X_{ij} - \bar{X}_{\bullet j \bullet})(X_{ij} - \bar{X}_{\bullet j \bullet})^t$ by replacing the null entries corresponding to each configuration; likewise, $E_{alt}$ is obtained from $\sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{ij} - \bar{X}_{\bullet jk})(X_{ij} - \bar{X}_{\bullet jk})^t$ in the same manner.

Under the null hypothesis, $-2 \ln \lambda$, follows asymptotically a Chi-square distribution. For small samples a permutation test may be used to approximate the distribution of this test statistic.

A simultaneous analysis of all the $Y$'s distributional-valued variables may be accomplished by a $p \times (m+1)$-dimensional MANOVA, following the same procedure.

In both cases, the covariance configuration is chosen by minimising the Bayesian Information Criterion (BIC) for the complete model.

## 5 Discriminant Analysis

Suppose now that the $n$ units under analysis belong to $K$ different populations, which are to be discriminated.

The model proposed above allows for parametric Discriminant Analysis of distributional data. For each configuration, an estimate of the optimum classification rule can be obtained with the corresponding $\Sigma$, by directly generalising the classical linear and quadratic discriminant classification rules. Let $G$ be the indicator of the predicted group, then we have:

**Linear classification rule:**

$$G = argmax_k \left( \hat{\mu_k}^t \hat{\Sigma}^{-1} X - \frac{1}{2} \hat{\mu_k}^t \hat{\Sigma}^{-1} \hat{\mu_k} + log\ \hat{\pi_k} \right)$$

**Quadratic classification rule:**

$$G = argmax_k \left( -\frac{1}{2} X^t \hat{\Sigma_k}^{-1} X + \hat{\mu_k}^t \hat{\Sigma_k}^{-1} X + log\ \hat{\pi_k} - \frac{1}{2} (log\ det \hat{\Sigma_k} + \hat{\mu_k}^t \hat{\Sigma_k}^{-1} \hat{\mu_k}) \right)$$

It is well known that quadratic classification rules often perform poorly in situations where the true covariances are indeed different because of the large number of parameters that need to be estimated (see e.g. McLachlan 1992). The restricted covariance configurations for the model proposed here allow for sparse classification rules, which are then less prone to this problem. Furthermore, in this way, sparsity is achieved in a natural manner, consistent with the nature of distributional data.

Since in real life applications the Normal distribution almost never holds true, there has always been considerable interest in finding how normality based classification

rules behave under different types of deviations from the model assumptions. Many studies (e.g. Ahmed and Lachenbruch 1977; Chinganda and Subrahmaniam 1979; Balakrishnan and Kocherlakota 1985; Duarte Silva et al. 2002; Rausch and Kelley 2009) have found the linear rule to be fairly robust to deviations from normality, although it can be adversely affected by severe skewness in the underlying distributions. However, Ashikaga and Chang (1981) found that "a more important issue than non-normality is whether the distributions of the two populations are similar in shape" (p. 680), and Nakanishi and Sato (1985) showed that, for fixed levels of skewness, "... [it] perform[s] best when the kurtosis are large" (p. 1190). The performance of the quadratic rule under non-normality was investigated, among others, by Clarke et al. (1979), Bayne and Tan (1981), Duarte Silva et al. (2002). These studies have found it to be mostly affected by severe skewness and small sample sizes. However, for large samples with symmetric, or mildly asymmetric, distributions it performed reasonably well.

## 6 Simulation study

### 6.1 Simulation design

To understand the performance of the discriminant methods based on the proposed model, we performed a controlled simulation experiment. The implemented procedure tries all five configurations in each case, and chooses the best one by minimising the BIC criterion.

We considered problems with three distributional variables. First we generated microdata, following a similar protocol to the one described in Dias et al. (2021), and detailed here below. Then we considered a full factorial design for problems with two and three groups, small and large samples, where groups may be balanced or imbalanced, and with different degrees of group separation, defined by similar/dissimilar means and/or standard deviations. Four alternatives are considered for the underlying distribution of the microdata, from being Uniform, Gaussian, to Log-Normal (a skewed distribution), or a mixture of those three.

For each data condition, 100 independent training samples were generated. The error rates of the classification rules were estimated by the average, across the 100 replicas, of the corresponding error proportions in an independently generated corresponding test sample.

The microdata has been generated according to the following specifications:

1. For each variable, the values of the mean and the standard deviation were fixed as: $\mu_{X1} = 20, \sigma_{X1} = 8; \mu_{X2} = 10, \sigma_{X2} = 6; \mu_{X3} = 5, \sigma_{X3} = 4$.
2. For each variable $j$ and for each group $k$, two vectors of length $n$ are generated, one with values of the means, $M_{X_{jk}} = [m_{jk}(i)]$ and another with values of the standard deviation $S_{X_{jk}} = [s_{jk}(i)]$. The $n$ values of each vector $M_{X_{jk}}$ and $S_{X_{jk}}$, are randomly generated, as follows:
   - $m_{jk}(i) \sim U(c_1(1+a), c_2(1+a))$, with $c_1 = 0.5 \times \mu_{Xj}, c_2 = 1.5 \times \mu_{Xj}, a = 0$ in Group 1 and $a > 0$ in the other groups.

- $s_{jk}(i) \sim U(h_1(1+b), h_2(1+b))$, with $h_1 = 0.5 \times \sigma_{Xj}$, $h_2 = 1.5 \times \sigma_{Xj}$, $b = 0$ in Group 1 and $b > 0$ in the other groups.

3. From each couple of values $m_{jk}(i)$ and $s_{jk}(i)$, $i \in 1, \ldots, n$ we randomly generate 10,000 real values, $x_{jki}(w)$, $w \in 1, \ldots, 10{,}000$ that allow obtaining the empirical distributions corresponding to unit $i$ and variable $Y_j$ in group $k$ and distribution $D$. According to the distribution $D$, the real values are generated as follows:
   - $D$ = Uniform distribution: $x_{jki}(w) \sim U(l_{jk}(i), u_{jk}(i))$ with $l_{jk}(i) = m_{jk}(i) - \sqrt{3}s_{jk}(i)$ and $u_{jk}(i) = m_{jk}(i) + \sqrt{3}s_{jk}(i)$
   - $D$ = Normal distribution: $x_{jki}(w) \sim N(m_{jk}(i), s_{jk}(i))$
   - $D$ = Log-Normal distribution: $x_{jki}(w) \sim LogLN(\tilde{m}_{jk}(i), \tilde{s}_{jk}(i))$ with $\tilde{m}_{jk}(i) = \frac{1}{2}\ln\left(\frac{m_{jk}(i)^4}{s_{jk}(i)^2 + m_{jk}(i)^2}\right)$ and $\tilde{s}_{jk}(i)) = \ln\left(\frac{s_{jk}(i))^2}{m_{jk}(i)^2}\right)$

4. From the 10,000 real values $x_{jki}(w)$ generated for each unit $i$, the following order statistics were computed: $Min$, $Q_1$ = 1st quartile, $Q_2$ = Median, $Q_3$ = 3rd quartile, $Max$. Each corresponding distribution is then represented by the Median, and the four LogRanges $\ln(Q_1 - Min)$, $\ln(Q_2 - Q_1)$, $\ln(Q_3 - Q_2)$, $\ln(Max - Q_3)$.

We then considered a full factorial design for problems with the following factors:

- Number of groups: two or three.
- Total sample size:
  - Training sample
    * Two groups: $n = 60$ (small sample), $n = 300$ (large sample) elements.
    * Three groups: $n = 90$ (small sample), $n = 300$ (large sample) elements.
  - Test sample
    * Two groups: $n = 1500$ elements.
    * Three groups: $n = 2700$ elements.
- Balance:
  - Training sample
    * Two groups: Equal sample sizes in the two groups; Unbalanced samples, with 1/3 elements in Group 1 and 2/3 in Group 2.
    * Three groups: Small sample: $n = 90 = 30+30+30$, or $n = 90 = 35+35+20$, or $n = 90 = 50 + 20 + 20$; large sample: $n = 300 = 100 + 100 + 100$, or $n = 300 = 120 + 120 + 60$, or $n = 300 = 180 + 60 + 60$.
  - Test sample
    * Two groups: Equal sample sizes in the two groups; Unbalanced samples, with 1/3 elements in Group 1 and 2/3 in Group 2.
    * Three groups: Equal sample sizes in the three groups; Unbalanced samples, with the same proportions as in the corresponding training samples.
- Group separation:
  - Case Ha: similar means and standard deviations in all groups
    * Two groups: $a = b = 0$ in Group 1, $a = b = 0.1$ in Group 2.

  - Three groups: $a = b = 0$ in Group 1, $a = b = 0.1$ in Group 2, $a = b = 0.2$ in Group 3.
- Case Hb: similar means and different standard deviations
  - Two groups: $a = b = 0$ in Group 1, $a = 0.1, b = 0.5$ in Group 2.
  - Three groups: $a = b = 0$ in Group 1, $a = 0.1, b = 0.5$ in Group 2, $a = 0.2, b = 1.0$ in Group 3.
- Case Hc: different means and similar standard deviations
  - Two groups: $a = b = 0$ in Group 1, $a = 0.5, b = 0.1$ in Group 2.
  - Three groups: $a = b = 0$ in Group 1, $a = 0.5, b = 0.1$ in Group 2, $a = 1.0, b = 0.2$ in Group 3.
- Case Hd: different means and different standard deviations
  - Two groups: $a = b = 0$ in Group 1, $a = 0.5, b = 0.5$ in Group 2.
  - Three groups: $a = b = 0$ in Group 1, $a = 0.5, b = 0.5$ in Group 2, $a = 1.0, b = 1.0$ in Group 3.

We note that in case Hb the standard deviations in Groups 2 and 3 are not only larger but also more disperse within each group (since multiplying the domain limits of a Uniform distribution by a given constant, in this case 1+b, not only impacts its location but also its variance). The same applies to the means in case Hc.

- Distribution of the microdata: Uniform, Normal, LogNormal, Mixture (uniform mixture of the previous three).

### 6.2 Discussion of results

The simulation results are summarized in Tables 13 to 28 in Supplementary Material. Figures 3 to 5 illustrate the obtained results, showing average error rates for Linear Discriminant Analysis (lda). The corresponding results for Quadratic Discriminant Analysis (qda) are in general clearly worse under the considered conditions, and for that reason are not displayed, and will not be analysed in detail. Quadratic Discriminant Analysis may however perform well under different data conditions, not covered in this simulation study.

We first note that error rates are similar across the distributions of the microdata, as it can be seen in Fig. 3, where all lines are basically horizontal. As such, we discuss graphical representations only for the Uniform distribution.

As expected, the error rates are smaller when both means and standard deviations are widely apart (condition Hd) and larger when they are both similar (case Ha), with the other two cases (cases Hb and Hc) leading to intermediate results—see Figs. 4 and 5.

In the two-group balanced setup, when the underlying distributions have similar means (cases Ha and Hb) the average error rates do not differ much across groups; when distributions have different means, the group 2 where those are larger and more disperse presents higher average error rates. In the three-group balanced setup, the "intermediate" group 2 (where mean values and standard deviations are between those of the other two groups) always displays higher global average error rates, as expected; however, error rates clearly differ across groups.

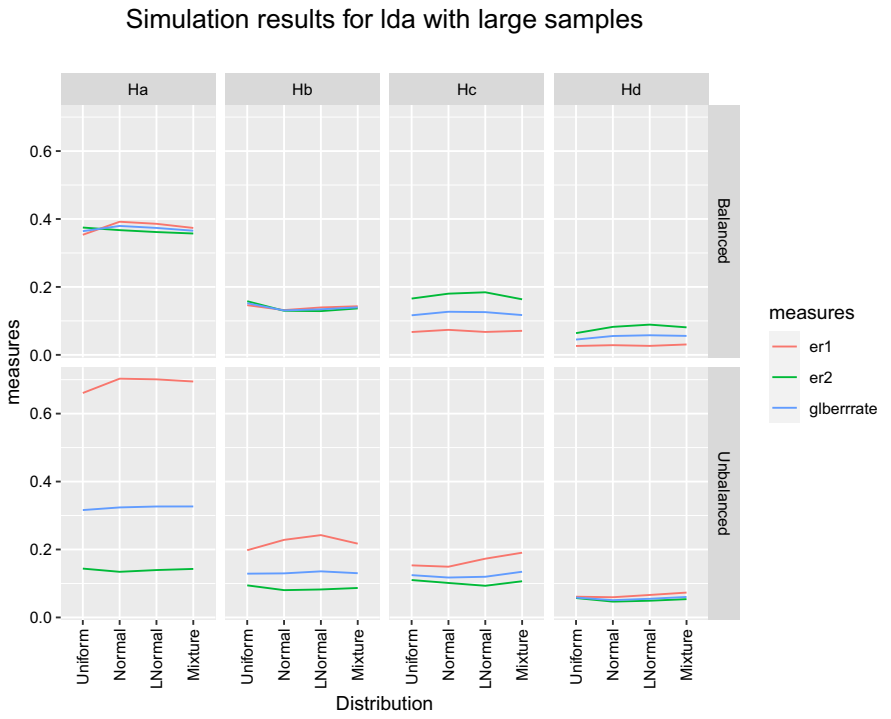


**Fig. 3** Error rates in the two group problem, by distribution, group separation, and balance

When the groups are unbalanced, average error rates tend to increase in smaller groups. In the two-group setup, the unbalance more than compensates other effects, and the small group 1 is more difficult to separate in all conditions. In the three-group setup, in the large–large–small condition, the small group 3 has clearly higher error rates; in the large–small–small condition the large group 1 has error rates inferior to those of the other two groups, and the "intermediate" group 2 is by far the most difficult, across all four data conditions (Ha to Hd).

Overall the lda method performed well, and in particular in the well separated case (Hd), average global error rates are below 8% in the two-group case, and below 16% in the three-group case.

## 6.3 Comparison with alternative approaches

### 6.3.1 Two-group problem

In the case of problems with two groups, we have compared the performance of the proposed approach with that of the method by Dias et al. (2021), as well as with interval parametric Discriminant Analysis as described in Duarte Silva and Brito (2015), using the simulation setup described in Dias et al. (2021).

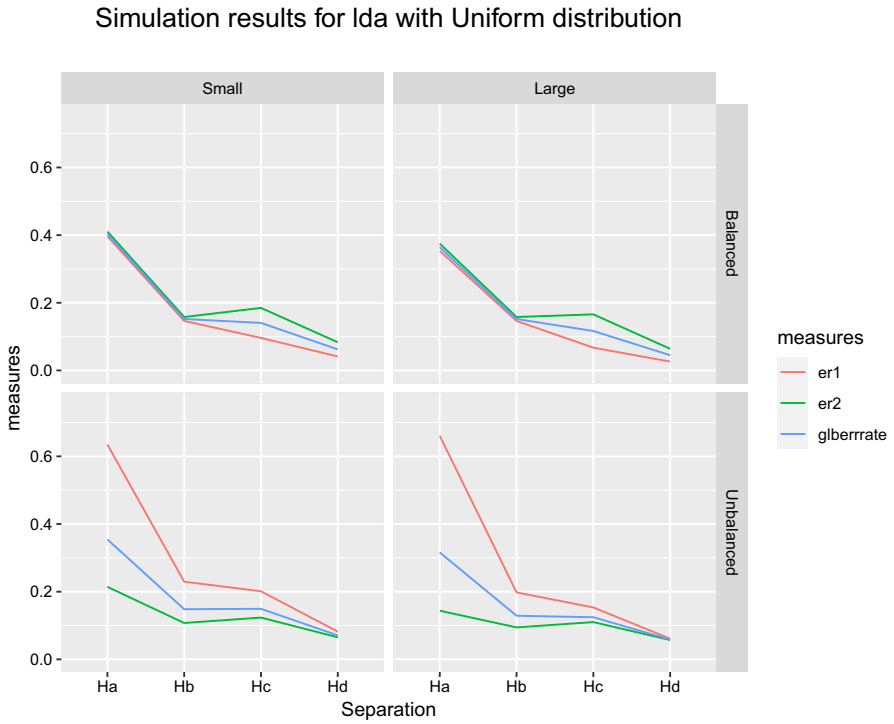


**Fig. 4** Error rates in the two group problem, by group separation, sample size, and balance

Again, we considered problems with three distributional variables. First we generated microdata, now exactly as in Dias et al. (2021), and as detailed here below. Then we considered a full factorial design with small and large samples, where groups may be balanced or imbalanced, with different degrees of group separation (defined by similar/dissimilar means and/or standard deviations) and considering four alternatives for the underlying distribution of the microdata: Uniform, Gaussian, Log-Normal, or a uniform mixture of those three.

For each data condition, 100 independent training samples were generated. The error rates of the classification rules were estimated by the average, across the 100 replicas, of the error proportions in a corresponding, independently generated, test sample.

The microdata has now been generated according to the following specifications:

1. For each variable, the values of the mean and the standard deviation were fixed as: $\mu_{X1} = 20, \sigma_{X1} = 8; \mu_{X2} = 10, \sigma_{X2} = 6; \mu_{X3} = 5, \sigma_{X3} = 4$.
2. For each variable $j$ and for each group $k$, two vectors of length $n$ are generated, one with values of the means, $M_{X_{jk}} = [m_{jk}(i)]$ and another with values of the standard deviation $S_{X_{jk}} = [s_{jk}(i)]$. The $n$ values of each vector $M_{X_{jk}}$ and $S_{X_{jk}}$, are randomly generated, as follows:

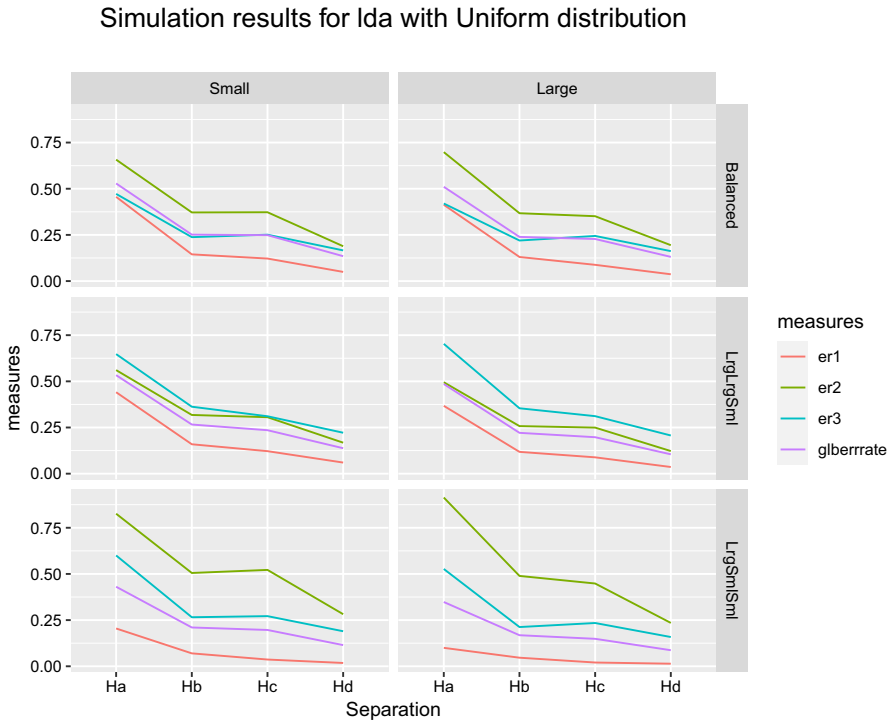


**Fig. 5** Error rates in the three group problem, by group separation, sample size, and balance

- $m_{jk}(i) \sim U(c_1(1+a), c_2(1+a))$, with $c_1 = 0.6 \times \mu_{Xj}$, $c_2 = 1.4 \times \mu_{Xj}$, $a = 0$ in Group 1 and $a > 0$ in the other groups.
- $s_{jk}(i) \sim U(h_1(1+b), h_2(1+b))$, with $h_1 = 0.6 \times \sigma_{Xj}$, $h_2 = 1.4 \times \sigma_{Xj}$, $b = 0$ in Group 1 and $b > 0$ in the other groups.

3. From each couple of values $m_{jk}(i)$ and $s_{jk}(i)$, $i \in 1, \ldots, n$ we randomly generate 5000 real values, $x_{jki}(w)$, $w \in 1, \ldots, 5000$ that allow obtaining the empirical distributions corresponding to unit $i$ and variable $X_j$ of the group $k$ and distribution $D$. According to the distribution $D$, the real values are generated as follows:

   - $D$ = Uniform distribution: $x_{jki}(w) \sim U(l_{jk}(i), u_{jk}(i))$ with $l_{jk}(i) = m_{jk}(i) - \sqrt{3}s_{jk}(i)$ and $u_{jk}(i) = m_{jk}(i) + \sqrt{3}s_{jk}(i)$
   - $D$ = Normal distribution: $x_{jki}(w) \sim N(m_{jk}(i), s_{jk}(i))$
   - $D$ = Log-Normal distribution: $x_{jki}(w) \sim LogLN(\tilde{m}_{jk}(i), \tilde{s}_{jk}(i))$ with $\tilde{m}_{jk}(i) = \frac{1}{2}\ln\left(\frac{m_{jk}(i)^4}{s_{jk}(i)^2 + m_{jk}(i)^2}\right)$ and $\tilde{s}_{jk}(i)) = \ln\left(\frac{s_{jk}(i))^2}{m_{jk}(i)^2}\right)$

4. From the 5000 real values $x_{jki}(w)$ generated for each unit $i$, the following order statistics were computed: $Min$, $Q_1$ =1st quartile, $Q_2$ = Median, $Q_3$ =3rd quartile, $Max$. Each corresponding distribution is then represented by the Median, and the four LogRanges $\ln(Q_1 - Min)$, $\ln(Q_2 - Q_1)$, $\ln(Q_3 - Q_2)$, $\ln(Max - Q_3)$.

The factorial design has the following factors:

- Total sample size:
  - Training sample: $n = 50$ (small sample), $n = 250$ (large sample) elements.
  - Test sample: $n = 1000$ elements.
- Balance: Equal sample sizes in the two groups; Unbalanced samples, with 1/5 elements in Group 1 and 4/5 in Group 2.
- Group separation:
  - Case Ha: similar means and standard deviations in both groups: $a = b = 0$ in Group 1, $a = b = 0.1$ in Group 2.
  - Case Hb: similar means and different standard deviations: $a = b = 0$ in Group 1, $a = 0.1, b = 0.5$ in Group 2.
  - Case Hc: different means and similar standard deviations: $a = b = 0$ in Group 1, $a = 0.5, b = 0.1$ in Group 2.
  - Case Hd: different means and different standard deviations: $a = b = 0$ in Group 1, $a = 0.5, b = 0.5$ in Group 2.

  The comments about cases Hb and Hc in the previous setup also apply here.
- Distribution of the microdata: Uniform, Normal, LogNormal, Mixture (uniform mixture of the previous three).

***Discussion of results***

Tables 29 to 32 in Supplementary Material gather the values of the average (and standard deviation) of the global error rates for all methods and data conditions considered. We note that in the case of small unbalanced samples, quadratic discriminant analysis with a distributional approach could not be applied since one of the groups does not contain enough observations to allow parameter estimation. Figures 6 and 7 display the average error rates for Linear Discriminant Analysis for interval and distributional approaches, as well as Dias et al. approach (Dias et al. 2021). Quadratic Discriminant Analysis provides generally higher error rates, and is hence not considered in these figures. As not all methods perform similarly across distributions, we

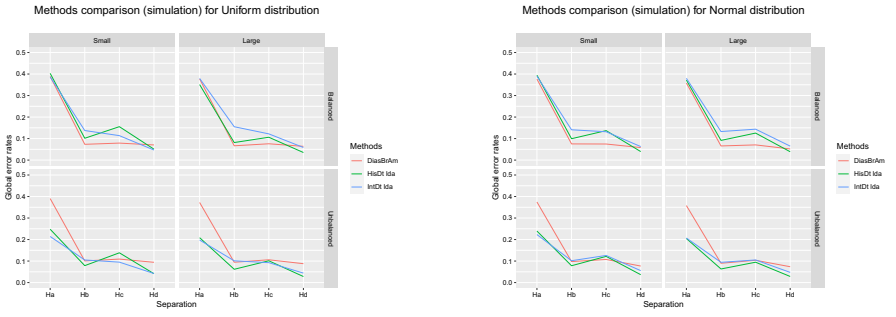


**Fig. 6** Comparison of error rates in the two group problem for different methods, by group separation and balance, Uniform (left) and Normal (right) distributions

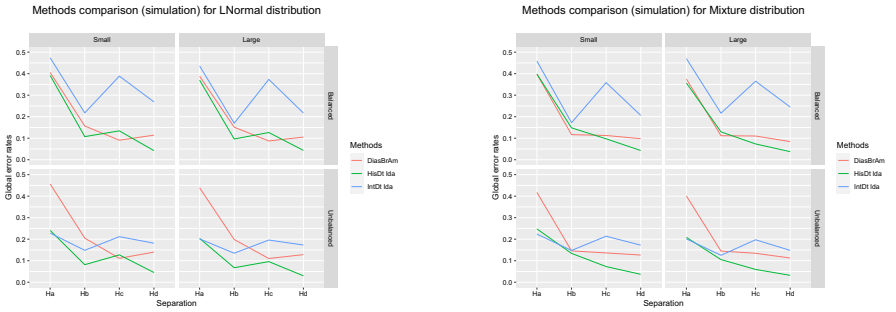


**Fig. 7** Comparison of error rates in the two group problem for different methods, by group separation and balance, LogNormal (left) and Mixture (right) distributions

discuss them all. The results for the Normal distribution follow closely those of the Uniform, and the results when a mixture of distributions is considered are frequently similar to those of the LogNormal distribution.

The proposed distributional approach tends to provide the best results when the groups are well separated (case Hd), particularly for unbalanced groups and when the underlying distributions are not symmetrical. In contrast, the method by Dias et al. performs better for the cases of intermediate separation (cases Hb and Hc), and balanced groups, with symmetrical distributions. In general, separation in terms of the standard deviations of the underlying distributions tends to favour the proposed approach, whereas separation in terms of the corresponding means tends to favour the approach by Dias et al. In particular, for case Hb with non-symmetrical distributions, or unbalanced groups, our approach (with lda) performs better; there is however a noticeably exception when Dias et al. approach provides lower average error rates when the groups are balanced and the data were generated by a mixture of distributions. In case Hc, Dias et al. approach performs better except for large unbalanced samples or microdata generated by a mixture of distributions.

When the groups are poorly separated (case Ha), and unbalanced, the parametric (interval and distributional) approaches perform much better, with similar results. Larger samples tend to favour the distributional approach, and smaller samples the interval approach, as expected. If however the groups are balanced then the interval approach provides higher error rates for non-symmetrical underlying distributions, whereas the proposed distributional approach and the method by Dias et al. perform similarly.

### 6.3.2 Three-group problem

We now compare the performance of the proposed approach in three-group problems with that of interval parametric Discriminant Analysis (Duarte Silva and Brito 2015), using the data generated in Sect. 6.1.

***Discussion of results***

Tables 33 to 36 in Supplementary Material gather the values of the average (and standard deviation) global error rates for all methods and data conditions considered;

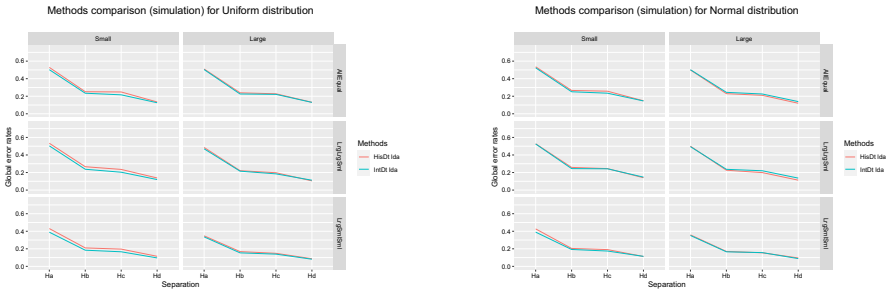


**Fig. 8** Comparison of error rates in the three group problem for distributional and interval approaches, by group separation and balance, Uniform (left) and Normal (right) distributions

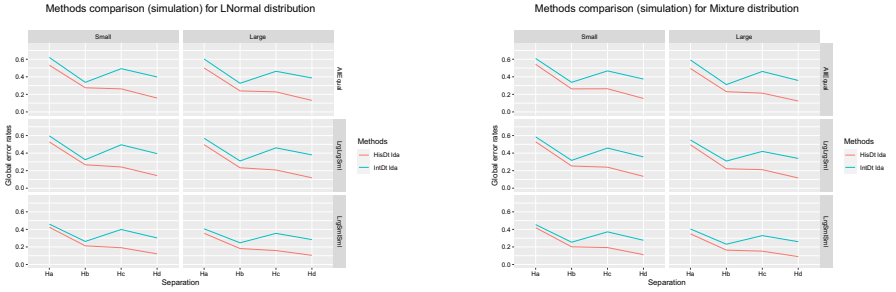


**Fig. 9** Comparison of error rates in the three group problem for distributional and interval approaches, by group separation and balance, LogNormal (left) and Mixture (right) distributions

Figs. 8 and 9 display the average rates for Linear Discriminant Analysis for interval and distributional approaches. As for the two-group problems, quadratic discriminant analysis provides higher error rates, and is not considered in these figures. When the underlying distributions are symmetrical both approaches provide similar results, with a slight advantage of the interval approach for the Uniform distribution. However, for non-symmetrical distributions, the proposed distributional approach is clearly superior. This highlights the advantage of the current proposal allowing taking into account the shape of the underlying distributions.

## 7 Applications

We now apply the proposed framework in two problems. The first one concerns Internet traffic, where it is wished to identify "attacks", leading to traffic re-direction. In the second one we analyse Portuguese official data, to identify area types (rural/medi–urban/urban) from the corresponding distributions of several variables connected to agricultural activity. Again, the implemented procedure tries all five configurations in each case, and chooses the best one now according to cross-validated estimates of the error rates. For comparison purposes, we also apply the parametric

Discriminant Analysis for interval-valued data, proposed in Duarte Silva and Brito (2015), the method developed in Dias et al. (2021), when applicable, and traditional Discriminant Analysis using only the medians or else the minima and the maxima of each variable.

### 7.1 Application to Internet traffic problem

We now consider the problem of identifying Internet traffic re-direction ("attacks") (Subtil et al. 2020). We use measurements obtained from a worldwide distributed probing platform, designed to detect routing variations based on round-trip-times (RTT) deviations inferred from disperse locations, the *targets*. There are 12 *probes* (see Table 3), where the RTT are measured, and four *targets*: Chicago, Frankfurt, Hong-Kong and London. Regular traffic goes from *probe* to *target* and comes back to the *probe*; in the presence of an attack, traffic is diverted through a *relay* before returning to the *probe*, and the RTT will be larger—see Table 3.

We note that Chicago and Frankfurt are both *targets* and *probes*, so that for those *targets* only 11 *probes* are considered.

The objective here is to determine if a *target* is under attack. For each *probe*, an observation consists of 10 measurements of RTT, and will hence be represented by distribution valued variables. We choose to characterize the corresponding empirical distribution by the median, and the intervals [minimum, median] and [median, maximum]. Table 4 shows the first and last units for the Chicago *target*. Table 5 presents the number of regular and diverted observations for each *target*.

In our model, $p = 12$ variables (the *probes*), $m = 2$ intervals, $C = Med$; $R_1^* = \ln(Med - Min)$; $R_2^* = \ln(Max - Med)$. Following the modelling above, we assume $(C, R_1^*, R_2^*) \sim N(\mu, \Sigma)$, where $\mu$ is a 36-dimensional vector and $\Sigma$ is the $36 \times 36$ covariance matrix.

**Table 3** Internet data: *Probes*, *targets* and *relays*

| | |
|---|---|
| *Probes* | |
| Amsterdam | Chicago |
| Frankfurt | Los Angeles 2 |
| Iceland | São Paulo |
| Milan | Viña del Mar |
| Sweden | Johannesburg 1 |
| Israel | Johannesburg 2 |
| *Targets* | |
| Chicago | Frankfurt |
| Hong Kong | London |
| *Relays* | |
| Los Angeles 1 | Madrid |
| Moscow | São Paulo 1 |

**Table 4** Internet data: distributional data for Chicago (partial view)

| | Probe: Amsterdam | ... | Probe: Sweden |
|---|---|---|---|
| $s_1$ | {[109.4, 115.7[, 0.5; [115.7, 117.7], 0.5} | ... | {[115.2, 118.7[, 0.5; [118.7, 135.0], 0.5} |
| $s_2$ | {[101.7, 117.6[, 0.5; [117.6, 121.3], 0.5} | ... | {[119.6, 122.0[, 0.5; [122.0, 124.3], 0.5} |
| ... | ... | ... | ... |
| $s_{11860}$ | {[103.1, 105.1[, 0.5; [105.1, 107.0], 0.5} | ... | {[118.4, 122.7[, 0.5; [122.7, 123.0], 0.5} |

**Table 5** Internet data: number of observations corresponding to regular traffic and traffic diverted by *relay*

| Target | Regular | Relay | | | | Total attacks |
|---|---|---|---|---|---|---|
| | | LA1 | Madrid | Moscow | São Paulo | |
| Chicago | 8732 | 848 | 598 | 768 | 914 | 3128 |
| Frankfurt | 8663 | 745 | 538 | 832 | 798 | 2913 |
| Hong Kong | 8544 | 770 | 564 | 808 | 810 | 2952 |
| London | 8569 | 681 | 567 | 782 | 835 | 2865 |

We shall address both a two-class (regular vs irregular traffic) and a five-class (regular, and four distinct *relays*) problems, for each *target*.

### 7.1.1 Analysis of Variance

We tested difference between the regular and irregular traffic, for all four *targets*, using all applicable *probes* as variables. The results in Table 6 show that those differences are clearly significant.

Univariate ANOVA's for each of the individual *probes* also provide highly significant results in all cases, with $p$-values always below $10^{-5}$.

We then considered the five-class problems, using the same approach. The results of the multivariate ANOVA's in Table 7 put in evidence the strong differences between the regular traffic and traffic diverted through each of the individual *relays*.

The univariate results for this case have $p$-values always below $10^{-30}$.

**Table 6** Internet data: multivariate Analysis of Variance results for the two-class problems

| Target | Configuration | Chi-square statistic | d.f | $p$-value |
|---|---|---|---|---|
| Chicago | 1 | 25,561.4 | 33 | $< 10^{-300}$ |
| Frankfurt | 1 | 11,850.8 | 33 | $< 10^{-300}$ |
| Hong Kong | 1 | 19,758.0 | 36 | $< 10^{-300}$ |
| London | 1 | 20,738.1 | 36 | $< 10^{-300}$ |

**Table 7** Internet data: multivariate Analysis of Variance results for the five-class problems

| Target | Configuration | Chi-square statistic | d.f | $p$-value |
|---|---|---|---|---|
| Chicago | 1 | 140,310.5 | 132 | $< 10^{-300}$ |
| Frankfurt | 1 | 114,739.1 | 132 | $< 10^{-300}$ |
| Hong Kong | 1 | 96,849.0 | 144 | $< 10^{-300}$ |
| London | 2 | 312,905.2 | 144 | $< 10^{-300}$ |

**N.B.:** Results for the individual ANOVA's are available from the authors upon request.

### 7.1.2 Discriminant Analysis

We have then applied Linear and Quadratic Discriminant Analysis (see Sect. 5) to this dataset, where we choose the covariance configuration by minimising a cross-validated estimate of the error rate. For comparison purposes, we have also performed classification on the classical dataset where for each unit and each *probe* only the median of the observed RTT values is kept, again using linear and quadratic classification rules.

In order to assess the extent of deviations from normality, Tables 60–63 and 64–71 in the Supplementary Material report the skewness and kurtosis coefficients for each variable/class combination in the two-class and five-class problems respectively. Even though a few variables show considerable skewness and excess kurtosis, high values in these two measures tend to go hand in hand, ruling out the most unfavourable data condition for the use of normality based linear classification rules (Nakanishi and Sato 1985). Furthermore, as it will be discussed below, although these rules are not optimal for these data (because of their lack of normality), they are actually quite effective, leading to very small error rates.

Tables 8 and 9 show the error rates for the best configuration with the proposed approach, the best configuration with interval Discriminant Analysis (as in Duarte Silva and Brito (2015)), traditional Discriminant Analysis on the medians, and the method proposed in Dias et al. (2021) (for the two-class problem). Error rates are obtained by 10-fold cross validation replicated 20 times in all cases, except for the computationally heavier method in Dias et al. (2021) where they are obtained by 10-fold cross validation replicated five times. Complete tables with error rates, as well as confusion matrices for the best cases, are provided as Supplementary Material.

We notice that the global error rates in the best cases (see Tables 8 and 9) are generally low for all approaches, always below 5.5% in the two-class problems and below 1.5% in the five-class problems. Nevertheless, some attacks appear to be more difficult to detect than others. This is particularly the case when the *target* is Hong Kong, where attacks are much harder to identify. In the two-class problems, and except for the Hong Kong *target*, the distributional approach is always better at detecting attacks, although the results are more mixed as concerns the global error rates. In the five-class problem the error rates tend to be lower. For the London an Hong

**Table 8** Internet data: error rates for the two class problems: best cases

| Target | Approach | Method | Config | Regular | Attack | Global |
|---|---|---|---|---|---|---|
| Chicago | Distributional | Linear | 3 | 0.00013 | 0.00106 | 0.00038 |
| | Interval | Linear | 1 | 0.00015 | 0.00353 | 0.00107 |
| | Traditional-Median | Linear | – | 0.00105 | 0.00206 | 0.00132 |
| | Dias et al. | – | – | 0.00034 | 0.00156 | 0.00067 |
| Frankfurt | Distributional | Quadratic | 5 | 0.01774 | 0.00000 | 0.01233 |
| | Interval | Linear | 3 | 0.00569 | 0.15899 | 0.05248 |
| | Traditional-Median | Quadratic | – | 0.01460 | 0.00009 | 0.01095 |
| | Dias et al. | – | – | 0.00616 | 0.15814 | 0.05255 |
| Hong Kong | Distributional | Linear | 1 | 0.00233 | 0.11133 | 0.02931 |
| | Interval | Linear | 1 | 0.00193 | 0.12409 | 0.03208 |
| | Traditional-Median | Quadratic | – | 0.00863 | 0.03196 | 0.01462 |
| | Dias et al. | – | – | 0.00218 | 0.21000 | 0.05364 |
| London | Distributional | Linear | 1 | 0.00097 | 0.00497 | 0.00218 |
| | Interval | Linear | 1 | 0.00157 | 0.00520 | 0.00267 |
| | Traditional-Median | Linear | – | 0.00083 | 0.00585 | 0.00209 |
| | Dias et al. | – | – | 0.00057 | 0.00732 | 0.00261 |

Kong *targets*, the traditional and distributional approaches have a similar performance, slightly better than the interval approach; however, for Chicago and Frankfurt the distributional approach gives results similar to the interval approach, better than the traditional one, and is particularly better in detecting attacks. In the two-class problems, the proposed distributional approach outperforms the method by Dias et al. (2021), confirming the simulation results for well-separated groups.

In the two-class problems, quadratic rules perform best for Frankfurt when using the distributional approach, and for both Frankfurt and Hong Kong when using a traditional approach. In the former case, we note that the sparse configuration 5 (diagonal covariance matrix) is selected, which can be explained by the large number of parameters involved. In the five-class problems, linear rules perform always better than quadratic ones, both for the distributional as well as interval and traditional approaches. In the distributional approach either configuration 1, or the sparser configurations 3 or 4, are selected, ruling out independence of the distributional variables (configuration 2).

The confusion matrix for the five-class problems concerning Hong Kong shows that most of the errors result from confusing the Madrid and the Moscow *relays*. We suspect that the extra RTT to Hong Kong should be quite similar when the *relay* is either Madrid or Moscow, which may explain the higher error rates in this case. For all the remaining cases, the best distributional method always succeeds in achieving error rates below 1%.

**Table 9** Internet data: error rates for the five class problems: best cases

| Target | Approach | Method | Config | Regular | LA1 | Madrid | Moscow | São Paulo | Global |
|---|---|---|---|---|---|---|---|---|---|
| Chicago | Distributional | Linear | 1 | 0.00000 | 0.00000 | 0.00009 | 0.00156 | 0.00117 | 0.00020 |
| | | Linear | 4 | 0.00001 | 0.00000 | 0.00189 | 0.00000 | 0.00117 | 0.00020 |
| | Interval | Linear | 1 & 3 | 0.00000 | 0.00000 | 0.00189 | 0.00156 | 0.00000 | 0.00019 |
| | Trad.-Med | Linear | – | 0.00000 | 0.00000 | 0.00167 | 0.00261 | 0.00109 | 0.00034 |
| Frankfurt | Distributional | Linear | 3 | 0.00036 | 0.00000 | 0.00009 | 0.00156 | 0.00117 | 0.00025 |
| | Interval | Linear | 1 | 0.00039 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00027 |
| | Trad.-Med | Linear | – | 0.00080 | 0.00134 | 0.00186 | 0.00000 | 0.00000 | 0.00077 |
| Hong Kong | Distributional | Linear | 3 | 0.00553 | 0.00122 | 0.07793 | 0.03340 | 0.00000 | 0.01018 |
| | Interval | Linear | 3 | 0.00517 | 0.00330 | 0.11887 | 0.05470 | 0.00000 | 0.01341 |
| | Trad.-Med | Linear | – | 0.00765 | 0.00065 | 0.06057 | 0.02537 | 0.00000 | 0.01049 |
| London | Distributional | Linear | 3 | 0.00002 | 0.00153 | 0.00000 | 0.00000 | 0.00000 | 0.00014 |
| | Interval | Linear | 4 | 0.00001 | 0.00000 | 0.00000 | 0.00000 | 0.00020 | 0.00002 |
| | Trad.-Med | Linear | – | 0.00000 | 0.00147 | 0.00000 | 0.00000 | 0.00000 | 0.00009 |

### 7.2 Application to official data

In this application we analyse data from the Portuguese Agriculture Census of 2009. The data units are parishes, in each of which farms have been observed and the corresponding data collected. In our analysis, each parish is described by the distribution, across the corresponding farms, of five variables, namely, Utilised Agricultural Area (UAA), Total Grain Cereals (TGC), Total Meadows and Fodder Crops (TMFC), Total Arable Land (TAL), and Total of Permanent Crops (TPC). The original data comprised 714 parishes with 22 variables, but only 149 parishes had no missing values, one of which had an obviously incorrect value.

The final dataset comprises hence 148 parishes, of which 49 correspond to Predominantly Rural Areas (PRA), 58 to Medi-Urban Areas (MUA), and 41 to Predominantly Urban Areas (PUA). The number of farms per parish varies between 49 and 4104, with an average of 759.9. Per area type, these values are: PRA - 49 to 4104 (average = 1110.2); MUA - 84 to 2295 (average = 662.9), PUA - 66 to 1071 (average = 380.4). The problem consists in distinguishing the three groups of parishes, according to the area type.

Information about the Census and the questionnaire used may be found in Statistics Portugal (2009).

In our model, $p = 5, m = 4, C = Med; R_1^* = \ln(Q_1 - Min); R_2^* = \ln(Med - Q_1); R_3^* = \ln(Q_3 - Med); R_4^* = \ln(Max - Q_3)$. Following the modelling above, we assume $(C, R_1^*, R_2^*, R_3^*, R_4^*) \sim N(\mu, \Sigma)$, where $\mu$ is a 25-dimensional vector and $\Sigma$ is the $25 \times 25$ covariance matrix.

#### 7.2.1 Analysis of Variance

We conducted a global multivariate Analysis of Variance, as well as variable-wise Analysis of Variance for each distributional variable. The results are shown in Table 10.

We observe that the three groups are well separated by the overall set of variables. However, at 5% significance level, variables Total Grain Cereals (TGC) and Total Meadows and Fodder Crops (TMFC) do not distinguish the three area groups.

**Table 10** Official data: Analysis of Variance results

| Variable | Configuration | Chi-square statistic | d.f | $p$-value |
|---|---|---|---|---|
| Global | 1 | 133.70 | 50 | $1.46 \times 10^{-9}$ |
| UAA | 1 | 56.65 | 10 | $1.55 \times 10^{-8}$ |
| TGC | 1 | 14.12 | 10 | 0.167 |
| TMFC | 1 | 16.59 | 10 | 0.084 |
| TAL | 1 | 18.80 | 10 | 0.043 |
| TPC | 1 | 25.92 | 10 | 0.004 |

**Table 11** Official data: error rates, best cases

| Approach | Method | Config | PRA | MUA | PUA | Global |
|---|---|---|---|---|---|---|
| Distributional | Linear | 4 | 0.46552 | 0.32653 | 0.48780 | 0.42568 |
| Interval | Quadratic | 1 | 0.18966 | 0.42857 | 0.92683 | 0.47297 |
| Classical - Med | Linear | – | 0.53448 | 0.34000 | 0.75610 | 0.53020 |
| Classical - Min-Max | Linear | – | 0.44828 | 0.24490 | 0.68293 | 0.44595 |

### 7.2.2 Discriminant Analysis

We have applied Linear and Quadratic Discriminant Analysis (see Sect. 5) to these data choosing the covariance configuration by minimising a cross-validated estimate of the error rate. For comparison purposes, we have also performed classification using Discriminant Analysis for interval-valued data, as proposed in Duarte Silva and Brito (2015), as well as on the classical dataset where for each unit and each variable only the median or else the minimum and the maximum of the observed values are kept, in all cases using both linear and quadratic classification rules.

In order to assess the extent of deviations from normality, Table 75 in the Supplementary Material reports the skewness and kurtosis coefficients for each variable in each area type. A few variables (in particular the median indicators) show some skewness and excess kurtosis. However, for most of them these measures are similar across area type which suggest that the variable distributions tend to have similar shapes across groups. Furthermore, for the few variables where that is not the case, a high value for the skewness tends to be associated with a corresponding high value for the kurtosis coefficient.

Table 11 shows the error rates obtained by leave-one-out, for the best method and configuration (in the distributional and interval approaches), and for the best method when a traditional Discriminant Analysis is applied. Complete tables with error rates are provided as Supplementary Material.

The linear classification rule is best both for the distributional and the traditional approaches, with the sparse configuration 4—only the corresponding indicators may be correlated—being chosen in the distributional approach. However, when using an interval approach, Quadratic Discriminant Analysis based on unconstrained variance-covariance matrices provides the lowest global error rate.

This is a rather more difficult problem than the previous one, with high error rates for all approaches. Nevertheless, the distributional approach clearly improves the traditional one, with particular emphasis on the classification of the Predominantly Urban Areas (PUA). Notice that the error rate obtained when using the linear distributional approach is never above 50%, neither globally nor in the PUA class, which is not the case for any of the remaining approaches.

Table 12 shows the confusion matrix for the best case, i.e., linear distributional approach, with configuration 4. We notice that the two area types more difficult to distinguish are the Medi-Urban Areas (MUA) and the Predominantly Urban Areas (PUA).

**Table 12** Official data: confusion matrix, configuration 4, linear classifier (original classes in rows, predicted in columns)

| | PRA | MUA | PUA |
|---|---|---|---|
| PRA | 0.534 | 0.276 | 0.190 |
| MUA | 0.122 | 0.673 | 0.204 |
| PUA | 0.049 | 0.439 | 0.512 |

As it could be expected, the Predominantly Rural Areas (PRA) and the Predominantly Urban Areas (PUA) are relatively easily separated.

## 8 Conclusion

In this paper, we propose parametric models specific for numerical distributional-valued variables which allow for model-based multivariate analysis of distributional data. We have addressed Analysis of Variance and Discriminant Analysis using linear and quadratic classification rules. A simulation study analyses the effect of relevant factors in classification problems. Comparisons with alternative approaches show that our proposal is particularly advantageous in problems with unbalanced groups, non-symmetrical underlying distributions, or well-separated groups. Experimental results in two applications from different areas, and with different sizes and group separation, show the pertinence and usefulness of the proposed framework.

The approach presented here may be extended in a number of different ways. On the one hand, robust estimation and (distributional) outlier detection should be developed. Other multivariate methodologies such as model-based clustering may also be tackled. On the other hand, alternative distributions may be considered, e.g., the Skew-Normal distribution, to extend the proposed approach beyond the normality assumption. Furthermore, an R Package implementing our models and methods is currently under development.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11634-025-00624-x.

**Acknowledgements** This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within projects LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020 and UID/GES/00731/2019. The authors wish to thank Sónia Dias and Paula Amaral for their precious collaboration in the experimental analysis regarding the comparison with the method proposed in Dias et al. (2021). The authors also thank the anonymous reviewers for their constructive comments and suggestions, that helped improving the quality of the manuscript.

**Funding** Open access funding provided by FCT|FCCN (b-on).

## References

Ahmed SW, Lachenbruch PA (1977) Discriminant analysis when scale contamination is present in the initial sample. In: Van Ryzin J (ed) Classification and clustering. Academic Press, University of Wisconsin-Madison, Madison, pp 331–353

Arroyo J, Maté C (2009) Forecasting histogram time series with k-nearest neighbours methods. Int J Forecast 25(1):192–207

Arroyo J, González-Rivera G, Maté C, San Roque AM (2011) Smoothing methods for histogram-valued time series: an application to value-at-risk. Stat Anal Data Mining ASA Data Sci J 4(2):216–228

Ashikaga T, Chang P (1981) Robustness of Fisher's linear discriminant function under two-component mixed normal models. J Am Stat Assoc 76(375):676–680

Balakrishnan N, Kocherlakota S (1985) Robustness to nonnormality of the linear discriminant function: mixtures of normal distributions. Commun Stat Theory Methods 14(2):465–478

Bayne C, Tan W (1981) QDF misclassification probabilities for known population parameters. Commun Stat Theory Methods 10(22):2315–2326

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487

Billard L, Diday E (2006) Symbolic Data Analysis: conceptual statistics and data mining. Wiley, Chichester

Billard L, Kim J (2017) Hierarchical clustering for histogram data. Wiley Interdiscip Rev Comput Stat 9(5):1405

Bock H-H, Diday E (2000) Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, Berlin

Brito PM, Chavent M (2012) Divisive monothetic clustering for interval and histogram-valued data. In: ICPRAM 2012-1st international conference on pattern recognition applications and methods, pp 229–234

Brito P (2014) Symbolic Data Analysis: another look at the interaction of data mining and statistics. Wiley Interdiscip Rev Data Mining Knowl Discov 4(4):281–295

Brito P, Dias S (2022) Analysis of distributional data. CRC Press, Boca Raton

Brito P, Duarte Silva AP (2012) Modelling interval data with Normal and Skew-Normal distributions. J Appl Stat 39(1):3–20

Chinganda EF, Subrahmaniam K (1979) Robustness of the linear discriminant function to nonnormality: Johnson's system. J Stat Plan Inference 3(1):69–77

Clarke WR, Lachenbruch PA, Broffitt B (1979) How non-normality affects the quadratic discriminant function. Commun Stat Theory Methods 8(13):1285–1301

De Carvalho FdA, Balzanella A, Irpino A, Verde R (2021) Co-clustering algorithms for distributional data with automated variable weighting. Inf Sci 549:87–115

Delicado P (2011) Dimensionality reduction when data are density functions. Comput Stat Data Anal 55(1):401–420

Dias S, Brito P (2015) Linear regression model with histogram-valued variables. Stat Anal Data Min ASA Data Sci J 8(2):75–113

Dias S, Brito P, Amaral P (2021) Discriminant analysis of distributional data via fractional programming. Eur J Oper Res 294(1):206–218

Duarte Silva AP, Brito P (2015) Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. J Classif 32:516–541

Duarte Silva AP, Stam A, Neter J (2002) The effects of misclassification costs and skewed distributions in two-group classification. Commun Stat Simul Comput 31(3):401–423

Duarte Silva AP, Brito P, Filzmoser P, Dias JG (2021) MAINT. Data: modelling and analysing interval data in R. R J 13(2):336–364

Gonzalez-Rivera G, Arroyo J (2012) Time series modeling of histogram-valued data: the daily histogram time series of S &P500 intradaily returns. Int J Forecast 28(1):20–33

Grygar TM, Radojičić U, Pavlů I, Greven S, Nešlehová JG, Tůmová Š, Hron K (2024) Exploratory functional data analysis of multivariate densities for the identification of agricultural soil contamination by risk elements. J Geochem Explor 259:107416
Ichino M (2011) The quantile method for symbolic principal component analysis. Stat Anal Data Mining ASA Data Sci J 4(2):184–198
Ichino M (2022) The lookup table regression model for histogram-valued symbolic data. Stats 5(4):1271–1293
Ichino M, Umbleja K, Yaguchi H (2021) Unsupervised feature selection for histogram-valued symbolic data using hierarchical conceptual clustering. Stats 4(2):359–384
Irpino A, Verde R (2015) Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance. Adv Data Anal Classif 9:81–106
Irpino A, Verde R, De Carvalho FdA (2014) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Syst Appl 41(7):3351–3366
Irpino A, Verde R, Carvalho FdA (2017) Fuzzy clustering of distributional data with automatic weighting of variable components. Inf Sci 406:248–268
Irpino A, Verde R (2006) A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Data science and classification. Springer, Berlin, Heidelberg, pp 185–192
Jin H, Billard L (2022) Copulas and histogram-valued data. J Comput Graph Stat 32:1–28
Kim J, Billard L (2011) A polythetic clustering process and cluster validity indexes for histogram-valued objects. Comput Stat Data Anal 55(7):2250–2262
McLachlan G (1992) Discriminant analysis and statistical pattern recognition. Wiley, Chichester
Nakanishi H, Sato Y (1985) The performance of the linear and quadratic discriminant functions for three types of non-normal distribution. Commun Stat Theory Methods 14(5):1181–1200
Panaretos VM, Zemel Y (2019) Statistical aspects of Wasserstein distances. Annu Rev Stat Appl 6(1):405–431
Petersen A, Müller H-G (2016) Functional data analysis for density functions by transformation to a Hilbert space
Rausch JR, Kelley K (2009) A comparison of linear and mixture models for discriminant analysis under nonnormality. Behav Res Methods 41(1):85–98
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Seber GAF (1984) Multivariate observations. Wiley, Chichester
Statistics Portugal (2009) Recenceamento Agrícola 2009. http://ra09.ine.pt/xportal/xmain?xpid=RA2009&xpgid=ine_ra_sabermais. Accessed: 6 Dec 2024
Subtil A, Oliveira MR, Valadas R, Pacheco A, Salvador P (2020) Detecting internet-scale traffic redirection attacks using latent class models. In: Madureira AM, Abraham A, Gandhi N, Silva C, Antunes M (eds) Proceedings of the tenth international conference on soft computing and pattern recognition (SoCPaR 2018). Springer, Berlin, pp 370–380
Umbleja K, Ichino M, Yaguchi H (2021) Hierarchical conceptual clustering based on quantile method for identifying microscopic details in distributional data. Adv Data Anal Classif 15:407–436
Verde R, Irpino A, Balzanella A (2015) Dimension reduction techniques for distributional symbolic data. IEEE Trans Cybern 46(2):344–355
Verde R, Irpino A (2007) Dynamic clustering of histogram data: using the right metric. In: Selected contributions in data analysis and classification. Springer, Berlin, pp 123–134
Zhao Q, Wang H, Lu S (2022) M-LDQ feature embedding and regression modeling for distribution-valued data. Inf Sci 609:121–152