dataSDA: datasets and basic statistics for symbolic data analysis in R

Po-Wei Chen[1], Chun-houh Chen[2], Han-Ming Wu[1*]

[1] *Department of Statistics, National Chengchi University, Taipei City, Taiwan, R.O.C*
[2] *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, R.O.C.*

September 9, 2025

Abstract

Traditional datasets represent each variable with a single value per observation. As data volume and heterogeneity increase, variables are more usefully represented by multivalued descriptors—intervals, histograms, and probability distributions—collectively termed symbolic data. Such representations retain distributional information and variability, improving analysis and interpretation. We introduce **dataSDA**, an R package that curates symbolic datasets across research domains and supports reading, writing, converting, and summarizing symbolic variables. **dataSDA** draws on the design of **RSDA** and **HistDAWass**, and extends them by enabling aggregation of conventional (single-valued) data into symbolic form. We utilized benchmark datasets within the **dataSDA** package to demonstrate and compare clustering, classification, and regression analyses in R. By combining a repository of ready-to-use datasets with utilities for data transformation and descriptive statistics, **dataSDA** aims to serve as a central resource for the collection and processing of symbolic data and to lower barriers to research in symbolic data analysis. The package is freely available on the Comprehensive R Archive Network (CRAN).
Keywords: benchmark dataset; dataset repository; descriptive data analysis; R package; symbolic data analysis;
Classification codes: 62-04

1 Introduction

In conventional data tables, we often analyze datasets with just single-value variables. But with the growth and complexity of data, our collections have expanded and diversified. To handle this type of data efficiently and retain its core information, we have moved beyond using only single values. We now use formats that capture multiple values, such as intervals, histograms, and probability distributions. This representation of data is termed

*Corresponding author. Email: wuhm@g.nccu.edu.tw

"symbolic data." The analysis of interval-valued data often serves as the foundational approach for analyzing other types of symbolic data, such as multi-valued, modal-valued, and modal multi-valued data. One source of interval data is the aggregation of large datasets. Aggregation allows the data to retain a manageable size while preserving as much of the original information as possible. Diday (1988) [23] introduced the concept of symbolic data analysis (SDA), and Billard and Diday (2003, 2007) [8, 9] provided an overview of the statistical methodologies for analyzing such data. By exploring data at higher descriptive levels, SDA facilitates a more comprehensive grasp of the data's distribution, characteristics, and variability, which are not amenable to conventional statistical methods [3, 61, 64].

Within the framework of SDA, symbolic variables are categorized into categorical and numerical scales, as illustrated in Fig 1. Categorical symbolic variables denote qualitative data and are subdivided into various types, depending on the nature of the categories they signify. Single-valued variables are those where each observation corresponds to one category or value, whereas multi-valued variables may associate multiple values or categories with a single observation. Specifically, the categorical multi-valued modal variable signifies categories associated with frequency or probability distributions. On the other hand, numerical symbolic variables exhibit a spectrum of configurations: single-valued variables depict observations with individual numerical values, and interval-valued variables incorporate a minimum and a maximum to account for imprecision or uncertainty.

The predominant focus in SDA literature is on interval-valued data. Modal variables represent scenarios where numerical observations reflect sets of probability distributions or frequencies, suitable for data marked by stochastic variations or ambiguity. Meanwhile, function variables serve to illustrate trends, relationships, or patterns among observed values. The core principle of SDA is to perceive data as groups of observations instead of individual entities. The key idea in SDA is based on the *reference concepts*, which represent higher-level abstractions of data. For example, a reference concept could be a "species" in ecological studies or a "customer segment" in marketing. These concepts allow researchers to analyze data at a more meaningful level of granularity, capturing variability and uncertainty within aggregated datasets. SDA complements traditional statistical methods by enabling the analysis of complex and aggregated data structures. This approach provides researchers with a suite of exploratory and inferential tools for data analysis, including clustering, classification, regression, and visualization, enriching the analytical possibilities in the study of such data.

A variety of R packages, specifically tailored for symbolic data analysis, are available on the Comprehensive R Archive Network (CRAN), as illustrated in Table 1. These packages are engineered to execute statistical and machine learning methodologies, with a primary focus on handling interval-valued and histogram-valued data, encompassing tasks such as dimensionality reduction, clustering, classification, and regression. However, the availability of SDA R packages is relatively limited compared to those in other research fields. Many SDA packages come equipped with few built-in datasets, each featuring distinct symbolic variables, thereby facilitating the exploration of a diverse array of datasets across various domains. This diversity stimulates the inception of innovative strategies and solutions for a multitude of problem types and fortifies collaborative research and empirical studies by

establishing a shared platform for data exchange. Two databases notably excel in fulfilling this role: the UCI Machine Learning Repository [27], and the UCR Time Series Classification Repository [18]. The former, initiated by the University of California, Irvine, has evolved into a prominent repository in the realms of machine learning and data mining. In contrast, the latter was conceived to provide a consolidated platform for time series classification datasets, enabling researchers to proficiently compare and benchmark their methodologies. Both of these repositories operate on web-based platforms and furnish well-documented descriptions and references for datasets. They offer various filtering options, allowing users to pinpoint datasets that align with their requirements. For more examples of database development in specific areas, see references in the literature such as [42], [15], and [12].

To address the current scattering and accessibility problems of symbolic datasets, we have created **dataSDA**, an R package. This package integrates a database specifically for symbolic datasets, with the goal of making data sharing easier for researchers in this field. **dataSDA** offers standardized datasets, acting as a benchmark for comparing different algorithms and methodologies. This promotes the creation of stronger and more efficient solutions. **dataSDA** houses a diverse collection of datasets from different domains, allowing researchers to select the ones that fit their needs best. It also provides detailed metadata and descriptions for each dataset, helping researchers understand the complexities of the data. Unlike the previously mentioned R packages that only supply the datasets, **dataSDA** goes a step further. It enhances functionalities, allowing the generation of symbolic data by aggregating conventional data and offering other data manipulation functions. This is built on the framework of widely-used symbolic data packages like **RSDA** and **HistDAWass**. We have also incorporated basic descriptive statistics such as mean, variance, and covariance for interval-valued and histogram-valued variables. This addition is a boon for users who wish to leverage R to develop new SDA methods. Additionally, we have added search and filter functionalities to assist users find relevant datasets easily. Given these enhancements and features, we are confident that **dataSDA** stands as a thorough and invaluable database. It's poised to be a dependable resource for the symbolic data analysis community, fostering transparency and reproducibility in research. Moreover, **dataSDA** provides real-world datasets, serving as a practical learning tool for students and professionals in machine learning, data analysis, and related fields.

This article is structured as follows: First, we present the design of the **dataSDA** package and examine its capabilities for manipulating both interval-valued and histogram-valued symbolic data, including comparisons with existing packages in Section 2. Next, we provide a comprehensive treatment of descriptive statistics for symbolic variables, covering both interval-valued and histogram-valued cases in Sections 3 and 4. We then analyze the strengths and limitations of the proposed **dataSDA** package in Section 5. Following this methodological foundation, Section 6 presents a benchmarking study comparing clustering, classification, and regression algorithms for symbolic datasets, with all implementations drawn from SDA-related packages. Finally, we conclude with a discussion of future developments in Section 7.

Table 1: The main features and built-in datasets provided by the various R packages available on CRAN for symbolic data.

(a) interval-valued data

IntervalQuestionStat (0.1.0): Tools to Deal with Interval-Valued Responses in Questionnaires

- Features: arithmetic operations, Cronbach's α, distance, transformation
- Built-in datasets: `lackinfo`

intkrige (0.1.0): A Numerical Implementation of Interval-Valued Kriging

- Features: distance, kriging
- Built-in datasets: `ohtemp`, `utsnow`, `utsnow_dtl`, `utsnow_dtl2`

iRegression (1.2.1): Regression Methods for Interval-Valued Variables

- Features: regression
- Built-in datasets: `Cardiological.CR`, `Cardiological.MinMax`, `soccer.bivar`

MAINT.Data (2.6.2): Model and Analyse Interval Data

- Features: clustering, discriminant analysis, LRT, MANOVA, mixture model estimation, MLE
- Built-in datasets: `Abalone`, `Cars`, `ChinaTemp`, `LoansbyPurpose_minmaxDt`, `LoansbyRiskLvs_minmaxDt`, `LoansbyRiskLvs_qntlDt`, `nycflights`

RSDA (3.0.9): R to Symbolic Data Analysis

- Features: descriptive statistics, distance, K-means, KNN, MCFA, PCA, regression, RF, SVM
- Built-in datasets: `abalone`, `Cardiological`, `cardiologicalv2`, `facedata`, `lynne1`, `oils`, `USCrime`, `uscrime_int`, `uscrime_intv2`, `VeterinaryData`

symbolicDA (0.7.1): Analysis of Symbolic Data

- Features: decision tree, distance, dynamical clustering, HINoV, KDA, MDS, PCA, RF, SOM.
- Built-in datasets: `cars`, `data_symbolic`

ggESDA (0.2.0): Exploratory Symbolic Data Analysis with **ggplot2**

- Features: descriptive statistics, PCA
- Built-in datasets: `AbaloneIdt`, `BLOOD`, `Cardiological2`, `Environment`, `facedata`, `mtcars.i`, `mushroom`, `oils`

(b) histogram-valued data

HistDat (0.2.0): Summary Statistics for Histogram/Count Data

- Features: descriptive statistics

HistDAWass (1.0.6): Histogram-Valued Data Analysis

- Features: descriptive statistics, Batch SOM, Fuzzy c-means, hierarchical clustering, K-means, PCA, regression, time series
- Built-in datasets: `Age_Pyramids_2014`, `Agronomique`, `BLOOD`, `BloodBRITO`, `China_Month`, `China_Seas`, `OzoneFull`, `OzoneH`

(c) polygonal-valued data

psda (1.4.0): Polygonal Symbolic Data Analysis

- Features: descriptive statistics, regression
- Built-in datasets: `longair`, `saeb2017`, `wnba2014`

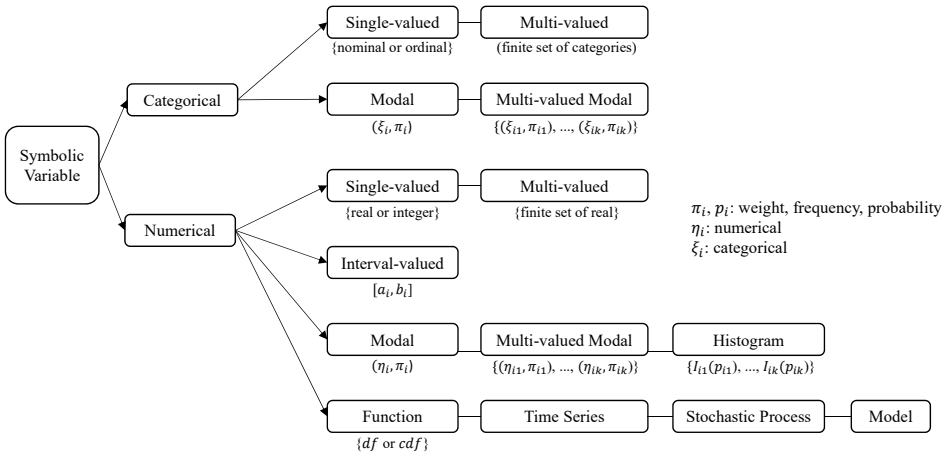

Figure 1: Types of symbolic variable diagram.

2 Methodology

2.1 The dataSDA package design

Fig 2 delineates the architectural design of the **dataSDA** package. When managing symbolic datasets in R, we contemplate two scenarios: the dataset to be imported is either conventional or symbolic. For external conventional data of varying file formats (e.g., `csv`, `xls`, etc.), it can be imported into R utilizing the appropriate R file read functions such as `read.csv` {`utils`} and `read_excel` {`readxl`}. For built-in conventional data, the dataset can be loaded directly into R using functions such as `load` {`base`} or `data` {`utils`}. Subsequently, various aggregation methods like K-means, hierarchical clustering, and user-defined methods are employed to aggregate the conventional data table into symbolic data table. If the external dataset to be imported is already in a symbolic format (e.g., `csv`, `sda`, `igap`, etc.), we have implemented several R functions to read these specific file formats. For accessing built-in symbolic datasets, users can employ the function `search_data` {`dataSDA`} to search or filter datasets that align with their requirements. See Section of the other functions in **dataSDA** for examples.

Once the imported dataset is classified as symbolic, data manipulations such as standardization can be performed, and datasets can be transformed from one symbolic class to another to align with the visualization and analysis methods provided by various SDA packages in R. **dataSDA** also furnishes descriptive statistics for interval-valued and histogram-valued data, including the mean, variance, and covariance of univariate and bivariate symbolic variables. Finally, the processed symbolic dataset can be exported to different file formats as needed. This streamlined process ensures that **dataSDA** is versatile and user-friendly, catering to a range of needs and preferences in the field of symbolic data analysis.

The **dataSDA** package incorporates the reference concept by providing tools to define, manipulate, and analyze symbolic data in the context of these higher-level abstractions. This approach enables researchers to gain deeper insights into complex datasets, making a base for modern data analysis. In its current state, the **dataSDA** package (version 0.1.0) incorporates 28 built-in datasets that are interval-valued. A summary of these datasets, including their subject area, analysis tasks, data size, and reference, is provided in Table 2. This package presently includes nine histogram-valued datasets, as outlined in Table 3.

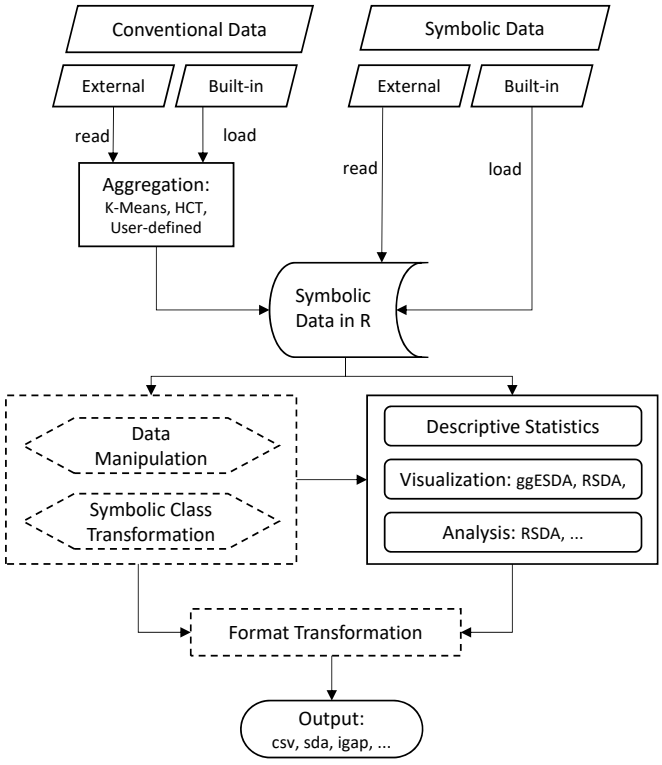

Figure 2: The design structure of the **dataSDA** package.

2.2 R functions for reading, writing, and converting symbolic data

While there are not an abundance of SDA-related R packages available on CRAN, the data formats and the classes defined for symbolic objects vary across each package. Additionally, the naming style for objects and R functions provided by these packages lacks consistency, which makes the code more challenging to use. Consequently, it is advantageous to unify most symbolic data objects into a same symbolic class for streamlined management. In this study, we primarily adopt features from the **RSDA** package to handle interval-valued data and the **HistDAWass** package to manage histogram-valued data, as these two packages

offer a broader range of analysis methods compared to others. Additionally, the naming convention for variables and functions within the **dataSDA** package follows the Tidyverse style guide [67].

There are several SDA-related packages dedicated to reading and writing, we mainly utilize the function `read.sym.table` provided by the **RSDA** package to import a symbolic data table from a `csv` file into R. The first row of the `csv` file should contain the variable names, prefixed with labels $C, $I, $H, or $S to identify their respective symbolic types. Here, $C denotes a continuous variable, $I indicates an interval variable, $H represents a histogram variable, and $S signifies a set variable. If an external dataset does not adhere to the pre-defined structure of a symbolic data table, we can also manually read the data into R and convert it to symbolic objects. For external conventional data tables, once read into R, we can employ `classic.to.sym` from **RSDA** to generate a symbolic data frame, specifying the type of symbolic data: interval, histogram, continuous, set, or modal.

Given a numerical data table, the **HistDAWass** package offers the `data2hist` function to read and generate a `distributionH` object, representing histogram-valued data. When input data are articulated through distributions (either empirical via histograms or theoretical through probability distributions), two functions, `distributionH` and `MatH`, provided by the **HistDAWass** package can be utilized. The `distributionH` function is used to create a histogram object associated with the `distributionH` class, while the `MatH` function is employed to create a matrix of histogram data (a MatH object) associated with the `MatH` class.

Both the **RSDA** and **dataSDA** packages offer functions, `write.sym.table` and `write_csv_table` respectively, to write symbolic data to a `csv` file. The distinction between these two functions resides in their output format. Additionally, the **symbolicDA** package implements the function `save.SO` to save a symbolic data table of the `symbolic` class to an `xml` file.

Since SDA-related packages have their own defined symbolic object formats and structures, the conversion among these formats is essential if one would like to utilize the corresponding SDA methods designed for that specified format. The **RSDA** package provides `SDS.to.RSDA` and `SODAS.to.RSDA` to convert the SDS SODAS and XML SODAS [24] formats to RSDA format. The **symbolicDA** package provides function `RSDA2SymbolicDA` to read a symbolic data table from a `csv` file or converts **RSDA** object to **SymbolicDA** "symbolic" class type object. The proposed **dataSDA** package enables more conversion functions between various formats of interval-valued symbolic objects, including SDS, SODAS, iGAP [41], and MM (minimum, maximum), as drawn in Fig 3. Note that, at present, there is no support for reverse conversions of SDS and SODAS formats.

To demonstrate the conversion of distinct classes of interval-valued datasets into the `symbolic_tbl` class, we use two representative examples: the `Abalone` and `mushroom` datasets. Due to space constraints, the detailed procedure and corresponding R code are provided in the package vignettes. These vignettes additionally demonstrate the conversion process for histogram-valued data into the `MatH` class from the **HistDAWass** package, illustrated with the `BLOOD` and `Weight` datasets.

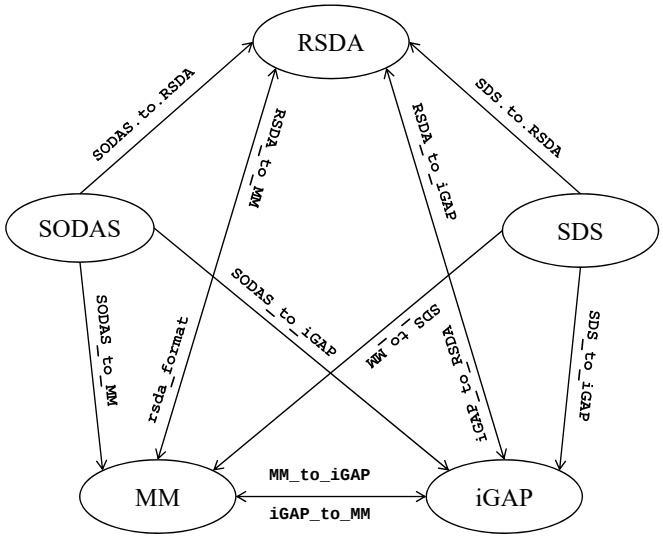

Figure 3: The R functions for interval-valued data format conversions in **dataSDA**.

2.3 Other functions in dataSDA

We have developed the `search_data` function to search and filter datasets that align with the specified options. The arguments for this function include `tag`, `task`, `n`, and `p`.

```
search_data(tag, task, type, n, p)
```

- `tag`: the source of dataset. For example, the datasets are sourced from the books such as Bock and Diday (2000) [10], Billard and Diday (2007) [5] and/or Diday and Noirhomme-Fraiture (2008) [24].
- `task`: regression, classification, clustering.
- `type`: interval, histogram, other.
- `n`: the number of observations in the dataset.
- `p`: the number of variables.

```
R> search_data(tag == "SDA_2007", task == "Regression",
+               type == "interval", n > 10, p < 10)
name subject_area       task     type  n p
1      Blood_pressure Life_Science Regression interval 15 3
2  Cardiological_data Life_Science Regression interval 11 3
3 Cardiological_data2 Life_Science Regression interval 15 3
4             Finance     Business Regression interval 14 7
```

```
reference      tag
1            Billard and Diday (2007) SDA_2007
2 Billard and Diday (2007), Xu (2010) SDA_2007
3 Billard and Diday (2007), Xu (2010) SDA_2007
4            Billard and Diday (2007) SDA_2007
```

2.4 Comparison with existing packages

The **dataSDA** package stands out from other R packages for symbolic data analysis (SDA) by offering three key advantages: (i) a comprehensive collection of built-in symbolic datasets, (ii) flexible conversion between multiple interval data formats, and (iii) an easy-to-use interface for computing various basic statistics using multiple formula options in one package. While some of the specialized SDA packages like **IntervalQuestionStat**, **intkrige**, **iRegression**, **MAINT.Data**, **RSDA**, **symbolicDA**, and **ggESDA** offer certain descriptive statistics and they focus exclusively on interval-valued data with rigid formatting requirements and limited datasets, **dataSDA** provides a more versatile solution. It supports not only interval-valued data but also histogram-valued and modal data. Its clear documentation and intuitive design make it accessible to beginners, unlike other packages that require advanced R skills. Currently, very few R packages support histogram-valued data (e.g., **HistDat**, **HistDAWass**) or other specialized types like polygonal-valued data (**psda**). While **dataSDA** is not designed for advanced analysis of these data types, it uniquely generate symbolic datasets by aggregating the conventional numerical tables based on the reference concept, enabling researchers to analyze data at a higher level of abstraction. This makes **dataSDA** particularly valuable for beginners for educational resources: instead of learning multiple specialized packages, users can begin exploring symbolic data and descriptive statistics with this single, comprehensive tool for handling complex and aggregated data structures.

3 Descriptive statistics for interval-valued data

In addition to amassing a collection of symbolic datasets and facilitating various format conversions in the proposed package, we also implement basic descriptive statistics such as mean, variance, covariance, and correlations for interval-valued and histogram-valued variables, drawing upon various formulas found in the literature. Descriptive statistics are pivotal in aiding our understanding and illustration of dataset characteristics, providing invaluable references for data analysis and statistical inference within research and decision-making processes. In this section, we outline the descriptive statistics and implement them as the functions `mean_int`, `var_int`, `cov_int`, and `cor_int` in the **dataSDA** package. We also list several R functions from different packages that are used for descriptive statistics of interval-valued data in Table 4.

Assume $\mathbf{X} = (X_1, \cdots, X_p)^T$ are p-dimensional numerical variables. Let $\{\mathbf{x}_i, i = 1, \cdots, n\}$ be the realizations of the variables $\mathbf{X}$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ is the ith numerical observation. Similar to the numerical case, we assume $\mathbf{\Xi} = (\Xi_1, \cdots, \Xi_p)^T$

are p-dimensional interval-valued variables. Denote the realizations of the interval-valued variables $\boldsymbol{\Xi}$ by $\{\boldsymbol{\xi}_i, i = 1, \cdots, n\}$, where $\boldsymbol{\xi}_i = (\xi_{i1}, \cdots, \xi_{ip})^T$, $\xi_{ij} = [a_{ij}, b_{ij}]$, $a_{ij} \leq b_{ij}, i = 1, \cdots, n, j = 1, \cdots, p$.

3.1 Quantification approaches

The quantification methods use appropriate coding to convert $\boldsymbol{\xi}_i$ into a form that can be processed using traditional calculations like the mean, variance, covariance, and correlations.

The center method (CM) The center method focuses on the midpoints of intervals as the key reference, and it does not take into account the range of variation within the data. Let $\boldsymbol{\xi}^c = (\boldsymbol{\xi}_1^c, \cdots, \boldsymbol{\xi}_n^c)^T$ represent the centers and $\boldsymbol{\xi}_{ij}^c$ be the ijth element of $\boldsymbol{\xi}^c$ where

$$\boldsymbol{\xi}_i^c = \left(\frac{a_{i1} + b_{i1}}{2}, \cdots, \frac{a_{ip} + b_{ip}}{2} \right)^T,$$

The matrix $\boldsymbol{\xi}^c$ is then handled as if it represents traditional data with p variables for n observations.

The vertices method (VM) The vertices method compiles the vertices of each hyperrectangle to create the vertices matrix. Let q_i be the number of nontrivial intervals for the ith observation. That is, $q_i = \sum_{j=1}^{p} I(a_{ij} < b_{ij})$, where $I(\cdot)$ is an indicator function. Denote the vertices matrix by $\boldsymbol{\xi^v} = (\boldsymbol{\xi}_{V_1}^v, \cdots, \boldsymbol{\xi}_{V_n}^v)^T$, where the rows of matrix $\boldsymbol{\xi}^v$ represent all the vertices $V_{k_i}, k_i = 1, \cdots, 2^{q_i}, i = 1, \cdots, n$, of n hyperrectangles and

$$\boldsymbol{\xi}_{V_i}^v = \begin{pmatrix} a_{i1} & \cdots & a_{ip} \\ a_{i1} & \cdots & b_{ip} \\ \vdots & \vdots & \vdots \\ b_{i1} & \cdots & a_{ip} \\ b_{i1} & \cdots & b_{ip} \end{pmatrix}^T, \quad i = 1, \cdots, n.$$

The dimension of the vertices matrix is determined by the sum of the vertices across the n hyperrectangles, multiplied by the number of variables, i.e., $\sum_{i=1}^{n} 2^{q_i} \times p$.

The quantile method (QM) Define the quantile matrix of $\{\boldsymbol{\xi}_i, i = 1, \cdots, n\}$ by $\boldsymbol{\xi}^q = (\boldsymbol{\xi}_{Q_1}^q, \cdots, \boldsymbol{\xi}_{Q_n}^q)^T$, where

$$\boldsymbol{\xi}_{Q_i}^q = \begin{pmatrix} q_{i01} & \cdots & q_{i0p} \\ \vdots & \vdots & \vdots \\ q_{im1} & \cdots & q_{imp} \end{pmatrix}^T, \text{ and } q_{ikj} = a_{ij} + (b_{ij} - a_{ij}) \frac{k}{m}$$

for $i = 1, \cdots, n, j = 1, \cdots, p, k = 0, \cdots, m$. The user determines the number $m (m \geq 1)$. It's important to note that building q_{ikj} assumes that each observed interval is uniform.

The dimensions of $\boldsymbol{\xi}^q$ are $(m+1) \times n$ by p. The key benefit of the quantile method is its ability to handle histograms, nominal multi-value types, and various other types of variables all at once.

The stacked endpoints method (SE) The stacked endpoints method stacks the minimum and maximum values of an interval to create the stacked matrix $\boldsymbol{\xi}^s$, in which the ith individual is

$$\boldsymbol{\xi}_i^s = \begin{pmatrix} a_{i1} & \cdots & a_{ip} \\ b_{i1} & \cdots & b_{ip} \end{pmatrix}, \quad i = 1, \cdots, n.$$

The stacked matrix is a specialized form of the QM with $m = 1$ and of size $2n \times p$.

The fitted values method (FV) The fitted values method draws inspiration from the MinMax method [55], which applies the MinMax approach to fit a linear regression model to symbolic interval-valued variables. The FV method quantifies interval variables using the fitted values from the maximum of the intervals, based on the simple linear regression model:

$$\hat{b}_{ij} = \hat{\eta}_{0j} + \hat{\eta}_{1j} a_{ij}, \quad j = 1, \cdots, p, \quad i = 1, \cdots, n.$$

The transformed matrix is denoted by $\boldsymbol{\xi}^f = (\boldsymbol{\xi}_1^f, \cdots, \boldsymbol{\xi}_n^f)^T$ where $\boldsymbol{\xi}_i^f = (\hat{b}_{i1}, \cdots, \hat{b}_{ip})^T$.

3.2 Distributional approaches

Distributional approaches for interval-valued data assume that the possible observations u_{ij} within a given interval $\xi_{ij} = [a_{ij}, b_{ij}]$ follow a uniform distribution over that interval. Additionally, it is assumed that each individual has equal probability of being observed. With these assumptions, Bertrand and Goupil (2000) [4] derived the empirical density for an interval variable Ξ_j as a mixture of the n uniform distributions,

$$f_j(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(u \in [a_{ij}, b_{ij}])}{b_{ij} - a_{ij}}, \quad j = 1, \cdots, p.$$

The symbolic sample mean and variance of Ξ_j can be expressed as follows:

$$\bar{\xi}_j = \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij}), \quad j = 1, \cdots, p,$$

and

$$\varsigma_j^2 = \frac{1}{3n} \sum_{i=1}^{n} (b_{ij}^2 + a_{ij} b_{ij} + a_{ij}^2) - \frac{1}{4n} \left[\sum_{i=1}^{n} (a_{ij} + b_{ij}) \right]^2. \tag{1}$$

Below, we introduce three methods for determining the symbolic sample covariance of interval-valued variables.

The empirical joint density method (EJD) The empirical joint density function for the interval variables Ξ_j and $\Xi_{j'}$ can be formulated as follows:

$$f_{jj'}(u,v) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(u \in [a_{ij}, b_{ij}], v \in [a_{ij'}, b_{ij'}])}{(b_{ij} - a_{ij})(b_{ij'} - a_{ij'})}, \quad j \neq j'.$$

Based on this empirical joint density function, Billard and Diday (2003) [8] derived the symbolic sample covariance for interval variables Ξ_j and $\Xi_{j'}$ as

$$\begin{aligned} \varsigma_{jj'} &= \frac{1}{4n}\sum_{i=1}^{n}[(a_{ij} + b_{ij})(a_{ij'} + b_{ij'})] \\ &\quad -\frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_{ij} + b_{ij})\right]\left[\sum_{i=1}^{n}(a_{ij'} + b_{ij'})\right]. \end{aligned} \tag{2}$$

Wang, Guan, and Wu (2012) [66] conceptualized the inner product for intervals, assuming that each data unit is a uniformly distributed random variable, densely populating the range $[a_{ij}, b_{ij}]$. Under these definitions and assumptions, they demonstrated that the sample variance of the centralized interval-valued variable Ξ_j and the sample covariance between two centralized interval-valued variables Ξ_j and $\Xi_{j'}$ can be precisely described by Eqs (1) and (2), excluding the latter parts of these equations. They also found that the off-diagonal elements of the variance-covariance matrices for $\boldsymbol{\xi}^c$ and $\boldsymbol{\xi}^v$ match exactly with those in $\varsigma_{jj'}$ from Eq (2). The sole distinction lies in the variances of the interval-valued variables.

The symbolic covariance method (GQ) Billard and Diday (2007) [5] reformulated the equation for the symbolic sample variance of Ξ_j in Eq (1) as

$$\varsigma_j^2 = \frac{1}{3n}\sum_{i=1}^{n}[(a_{ij} - \bar{\xi}_j)^2 + (a_{ij} - \bar{\xi}_j)(b_{ij} - \bar{\xi}_j) + (b_{ij} - \bar{\xi}_j)^2].$$

They expanded upon this equation to articulate the symbolic sample covariance for Ξ_j and $\Xi_{j'}$ as

$$\varsigma_{jj'} = \frac{1}{3n}\sum_{i=1}^{n} G_j G_{j'} \left[Q_j Q_{j'}\right]^{1/2}, \quad j, j' = 1, \cdots, p, \tag{3}$$

where for $J = j, j'$,

$$\begin{aligned} Q_J &= (a_{iJ} - \bar{\xi}_J)^2 + (a_{iJ} - \bar{\xi}_J)(b_{iJ} - \bar{\xi}_J) + (b_{iJ} - \bar{\xi}_J)^2, \\ G_J &= \begin{cases} -1, & \text{if } \xi_{iJ}^c \leq \bar{\xi}_j, \\ 1, & \text{if } \xi_{iJ}^c > \bar{\xi}_j, \end{cases} \end{aligned}$$

and ξ_{iJ}^c is the midpoint of the interval $[a_{iJ}, b_{iJ}]$.

The total sum of products (SPT) Billard (2007, 2008) [5, 6] further demonstrated that the sample variance in Eq (1) is a function of the total sum of squares (TSS) and that the TSS can be decomposed into the sum of the within variation and the between variation. The total sum of products (SPT) is the sum of the within sum of products and the between sum of products. Billard (2008) [6] extended Eq (1) to the bivariate case to obtain the sample covariance of Ξ_j and $\Xi_{j'}$ based on the decomposition of the SPT as follows.

Billard (2007, 2008) [5,6] further illustrated that the sample variance outlined in Eq (1) depends on the total sum of squares (TSS). He clarified that TSS can be decomposed into the sum of the within variations and the between variations. Therefore, the total sum of products (SPT) is likewise a combination of the within sum of products and the between sum of products. Billard (2008) [6] expanded this concept to two variables, calculating the sample covariance of Ξ_j and $\Xi_{j'}$ by similarly breaking down the SPT.

$$\begin{aligned} \varsigma_{jj'} &= \frac{1}{6n}\sum_{i=1}^{n}[2(a_{ij}-\bar{\xi}_j)(a_{ij'}-\bar{\xi}_{j'})+(a_{ij}-\bar{\xi}_j)(b_{ij'}-\bar{\xi}_{j'}) \\ &\quad +(b_{ij}-\bar{\xi}_j)(a_{ij'}-\bar{\xi}_{j'})+2(b_{ij}-\bar{\xi}_j)(b_{ij'}-\bar{\xi}_{j'})] \end{aligned} \tag{4}$$

The definitions and calculations of the symbolic sample covariance in Eqs (2)–(4) are consistent with the results in the classic data case if $a_{ij} = b_{ij}$ for $i = 1, \cdots, n, j = 1, \cdots, p$. If $j = j'$, Eq (4) reduces to the sample variance of the interval-valued variable j, as given in Eq (1).

3.3 An example

We use the functions `int_mean`, `int_var`, `int_cov`, and `int_cor` to calculate the mean, variance, covariance, and correlation of interval-valued variables. Table 5 presents the mean and variance of the variable `Pileus.Cap.Width`, as well as the covariance and correlation between `Pileus.Cap.Width` and `Stipe.Length` in the `mushroom` dataset, computed using various approaches. For comparison, we also compute the mean and variance of the variable `Sepal.Length`, along with the covariance and correlation between `Sepal.Length` and `Sepal.Width` in the `iris` dataset, applying the same set of approaches. The results are summarized in Table 6. In this example, each interval-valued variable in the `iris` dataset consists of identical left and right endpoints. It is notable that the results obtained using the QM and SE approaches for computing `var`, `cov`, and `cor` show slight differences compared to the other methods.

4 Descriptive statistics for histogram-valued data

Assume that the random variable of interest, Y, for observation ω_i, where $i = 1, \cdots, n$, takes values on the intervals $\xi_{ih} = [a_{ih}, b_{ih})$ with probabilities π_{ih}, for $h = 1, \cdots, H_i$. A typical example of such data occurs when an object ω_i is represented by a histogram as its observed value of Y, either alone or after aggregation. These data are known as modal interval-valued observations, also referred to as histogram-valued observations. Various R

functions have been developed to calculate descriptive statistics for histogram-valued data, as detailed in Table 7. These statistical measures have been extended to both univariate and bivariate histogram-valued variables, enhancing their analysis and interpretation.

4.1 Univariate descriptive statistics

Basic statistics according to Bertrand and Goupil By analogy with ordinary interval-valued variables, Bertrand and Goupil (2000) [4] assumed that within each interval $[a_{ih}, b_{ih})$ of a histogram observation ω_i, each individual description vector $x \in vir(d_i)$ is uniformly distributed across that interval. Therefore, for each ξ_h

$$P(x \leq \xi_h \,|\, x \in vir(d_i)) = \begin{cases} 0, & \xi_h < a_{ih}, \\ \dfrac{\xi_h - a_{ih}}{b_{ih} - a_{ih}}, & a_{ih} \leq \xi_h < b_{ih}, \\ 1, & \xi_h \geq b_{ih}. \end{cases}$$

For histogram-valued observations, the symbolic sample mean becomes

$$\bar{Y} = \frac{1}{2n} \sum_{i=1}^{n} \left[\sum_{h=1}^{H_i} (b_{ih} + a_{ih}) \pi_{ih} \right], \tag{5}$$

and the symbolic sample variance is

$$S^2 = \frac{1}{3n} \sum_{i=1}^{n} \left[\sum_{h=1}^{H_i} (b_{ih}^2 + b_{ih} a_{ih} + a_{ih}^2) \pi_{ih} \right] - \frac{1}{4n^2} \left[\sum_{i=1}^{n} \sum_{h=1}^{H_i} (b_{ih} + a_{ih}) \pi_{ih} \right]^2. \tag{6}$$

The mean and variance based on ℓ_2 Wasserstein distance Gilchrist (2000) [29] proposed a shift from traditional methods of computing descriptive statistics for histogram-valued variables, which typically assume univariate data with normal or near-normal distributions. This assumption falls short for histogram-valued variables due to their multivariate characteristics and potential non-normal distributions. It is crucial to embrace methods that account for the distribution's shape, location, and variability when dealing with such variables. The Wasserstein distance-based method stands out for its ability to introduce novel variability measures by assessing differences between distributions. This technique adeptly tackles the dual aspects of variability found in multivariate data, offering a more precise and thorough framework for summarizing and analyzing histogram-valued variables in comparison to conventional approaches.

Various formulations of the Wasserstein distance exist in the literature, we adopt the formalization presented by [58]. This reference also includes the principal sources on the Wasserstein metric. It defines the distance between two densities, ϕ_i and $\phi_{i'}$, using the quantile functions Φ_i^{-1} and $\Phi_{i'}^{-1}$ associated with their respective cumulative distribution functions (CDFs) Φ_i and $\Phi_{i'}$ as follows:

$$d_{W_p}(\phi_i, \phi_{i'}) \equiv \left(\int_0^1 \left| \Phi_i^{-1}(t) - \Phi_{i'}^{-1}(t) \right|^p dt \right)^{\frac{1}{p}}. \tag{7}$$

The formulation presented in [58] illustrates that the d_{W_p} distance can be viewed as an extension of the classical L_p Minkowski distance to quantile functions (QFs). According to [29], QFs possess several valuable statistical properties. Some of the most relevant for the discussions in this paper include: QFs being in a one-to-one correspondence with their corresponding density functions, QFs having a finite domain (with $t \in [0, 1]$), and QFs being non-decreasing functions. To simplify notation and avoid multiple indices, we use d_W to denote the ℓ_2 Wasserstein (L2W) distance between two probability distributions as follows:

$$d_W(\phi_i, \phi_{i'}) \equiv \sqrt{\int_0^1 \left[\Phi_i^{-1}(t) - \Phi_{i'}^{-1}(t)\right]^2 dt}. \tag{8}$$

In this case, the Fréchet mean of distribution variable Y, with respect to d_W (assuming equal weights w_i), is the density function, corresponding to the mean quantile function, that solves the following optimization problem:

$$\begin{aligned} M_W(Y) &= \arg\min_x \sum_{i=1}^{n} d_W^2(\phi_i, x) = \bar{\phi} \\ &= \frac{d(\bar{\Phi}^{-1})^{-1}}{dy} = \frac{d\bar{\Phi}}{dy}, \text{ where } \bar{\Phi}^{-1} = \frac{1}{n}\sum_{i=1}^{n} \Phi_i^{-1}. \end{aligned} \tag{9}$$

The mean of $M_W(Y)$ is the arithmetic mean of the means of the quantile functions:

$$\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \mu_i. \tag{10}$$

The variance of $M_W(Y)$ mean distribution function can be expressed by a function of the standard deviations of the single distributions and of the correlation terms between pairs of quantile functions, as follows:

$$\sigma_{\bar{y}}^2 = \sum_{i=1}^{n} \left[\frac{\sigma_i}{n}\right]^2 + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j>i} [\rho_{i,j}\sigma_i\sigma_j]. \tag{11}$$

Given $M_W(Y)$, the mean of a set of n units described by the distributional symbolic variable Y, the variance of Y can be defined as the mean of the squared ℓ_2 Wasserstein distance between each distribution $y(i)$ and $M_W(Y)$ as follows:

$$S_W^2(Y) = \left[\frac{1}{n}\sum_{i=1}^{n} \mu_i^2 - \mu_{\bar{y}}^2\right] + \left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{i,j}\sigma_i\sigma_j\right]. \tag{12}$$

4.2 Bivariate descriptive statistics

The symbolic covariance method according to Bertrand and Goupil (2000) Bertrand and Goupil (2000) [4] introduced a preliminary definition of covariance, denoted

as $C_{BG}(Y_1, Y_2)$, between two numeric-symbolic variables using the following formula:

$$C_{BG}(Y_1, Y_2) = \frac{1}{n}\sum_{i=1}^{n} \mu_{i1} \cdot \mu_{i2} - \bar{Y}_1\bar{Y}_2 \tag{13}$$

where $\bar{Y}_1 = \frac{1}{n}\sum_{i=1}^{n} \mu_{i1}$ and $\bar{Y}_2 = \frac{1}{n}\sum_{i=1}^{n} \mu_{i2}$. In this case, $C_{BG}(Y_1, Y_2)$ represents the covariance of the mean values of the distribution-valued data. A significant limitation of this formulation is that, if $Y_1(i) = Y_2(i)$ for each $i = 1, \cdots, n$, then $C_{BG}(Y_1, Y_2)$ does not equal $S^2(Y_1)$ or $S^2(Y_2)$.

The symbolic covariance method according to Billard and Diday (2007) Billard and Diday (2007) [5] proposed an explicit formulation for the covariance between histogram-valued variables, treating a pair of such variables as a weighted combination of intervals. Let Y_1 and Y_2 be two histogram-valued data sets:

$$Y_1(i) = \{([a_{i1,1}, b_{i1,1}], \pi_{i1,1}), \cdots, ([a_{i1,H_{i1}}, b_{i1,H_{i1}}], \pi_{i1,H_{i1}})\}$$

where $\sum_{h_1=1}^{H_{i1}} \pi_{i1,h_1} = 1$ and

$$Y_2(i) = \{([a_{i2,1}, b_{i2,1}], \pi_{i2,1}), \cdots, ([a_{i2,H_{i2}}, b_{i2,H_{i2}}], \pi_{i2,H_{i2}})\}$$

where $\sum_{h_2=1}^{H_{i2}} \pi_{i2,h_2} = 1$. The covariance for histogram-valued data is defined as:

$$C_{BD}(Y_1, Y_2) = \frac{1}{3n}\sum_{i=1}^{n}\sum_{h_1=1}^{H_{i1}}\sum_{h_2=1}^{H_{i2}} \pi_{i1,h_1}\pi_{i2,h_2} G_1 G_2 [Q_1 Q_2]^{1/2},$$

where

$$Q_j = (a_{ij,h_j} - \bar{Y}_j)^2 + (a_{ij,h_j} - \bar{Y}_j)(b_{ij,h_j} - \bar{Y}_j) + (b_{ij,h_j} - \bar{Y}_j)^2, \tag{14}$$

$$G_j = \begin{cases} -1 & \text{if } \frac{1}{2}\sum_{h_j=1}^{H_{ij}} \pi_{ij,h_j}(a_{ij,h_j} + b_{ij,h_j}) \le \bar{Y}_j, \\ 1 & \text{if } \frac{1}{2}\sum_{h_j=1}^{H_{ij}} \pi_{ij,h_j}(a_{ij,h_j} + b_{ij,h_j}) > \bar{Y}_j. \end{cases} \tag{15}$$

However, since $C_{BD}(Y_1, Y_2)$ is a reformation of $C_{BG}(Y_1, Y_2)$, it has the same limitation: if $Y_1(i) = Y_2(i)$ for each $i = 1, \cdots, n$, then $C_{BD}(Y_1, Y_2)$ is not equal to $S^2(Y_1)$ or $S^2(Y_2)$.

The symbolic covariance method according to Billard (2008)) In order to solve above drawback, Billard (2008) [6] proposed new covariance statistics for histogram-valued data as follows:

$$C_B(Y_1, Y_2) = \frac{1}{6n}\sum_{i=1}^{n}\sum_{h_1=1}^{H_{i1}}\sum_{h_2=1}^{H_{i2}} \left\{ \left[\begin{array}{l} 2(b_{i1,h_1} - \bar{X}_1)(b_{i2,h_2} - \bar{X}_2) \\ +(b_{i1,h_1} - \bar{X}_1)(a_{i2,h_2} - \bar{X}_2) \\ +(a_{i1,h_1} - \bar{X}_1)(b_{i2,h_2} - \bar{X}_2) \\ +2(a_{i1,h_1} - \bar{X}_1)(a_{i2,h_2} - \bar{X}_2) \end{array} \right] \pi_{i1,h_{i1}}\pi_{i2,h_{i2}} \right\}$$

However, in this case the same deficiency of $C_{BD}(Y_1, Y_2)$ arises. In fact, if $Y_1(i) = Y_2(i)$ for each $i = 1, \cdots, n$, then $C_B(Y_1, Y_2) \neq S^2(Y_1)$ or $C_B(Y_1, Y_2) \neq S^2(Y_2)$.

The symbolic covariance based on ℓ_2 Wasserstein distance The centrality property of $M_W(Y)$ is based on the correlation coefficient between two quantile functions, denoted as $\rho_{i,i'}$. It is defined as follows:

$$\rho_{i,i'} = \frac{\int_0^1 \left(\Phi_i^{-1}(t) - \mu_i\right)\left(\Phi_{i'}^{-1}(t) - \mu_{i'}\right) dt}{\sigma_i \sigma_{i'}} = \frac{\int_0^1 \Phi_i^{-1}(t)\Phi_{i'}^{-1}(t)\, dt - \mu_i\mu_{i'}}{\sigma_i\sigma_{i'}}. \tag{16}$$

For two quantile functions Φ_i^{-1} and Φ_j^{-1}, associated with two probability density functions (PDFs) ϕ_i and ϕ_j with means μ_i and μ_j and standard deviations σ_i and σ_j respectively, and in reference to Eq. (16), the product of two quantile functions is defined as:

$$\langle \Phi_i^{-1}, \Phi_j^{-1} \rangle = \int_0^1 \Phi_i^{-1}(t)\Phi_j^{-1}(t)\, dt = \rho_{i,j}\sigma_i\sigma_j + \mu_i\mu_j. \tag{17}$$

With the ℓ_2 Wasserstein metric and the associated product of QFs defined in Eq.(17), we propose an alternative approach for measuring the covariance and correlation between two symbolic variables, addressing some deficiencies of the approach by Billard and Diday (2007) [5]. Let Y_1 and Y_2 be two modal numeric variables describing a set of n units. The generic i-th unit is described by the ordered pair $y(i) = \{y_1(i), y_2(i)\}$, where $y_1(i) = \phi_{i1}$ and $y_2(i) = \phi_{i2}$ are density functions with respective means μ_{i1} and μ_{i2}, and standard deviations σ_{i1} and σ_{i2}. Associated with each ϕ_{i1} (resp. ϕ_{i2}) is the corresponding cumulative distribution function (CDF) Φ_{i1} (resp. Φ_{i2}) and the respective quantile function (QF) denoted Φ_{i1}^{-1} (resp. Φ_{i2}^{-1}). Let $C_W(Y_1, Y_2)$ be defined as the empirical covariance between Y_1 and Y_2 based on the ℓ_2 Wasserstein metric as follows:

$$C_W(Y_1, Y_2) = \frac{1}{n}\sum_{i=1}^{n} \int_0^1 \left[\Phi_{i1}^{-1}(t) - \bar{\Phi}_1^{-1}(t)\right] \cdot \left[\Phi_{i2}^{-1}(t) - \bar{\Phi}_2^{-1}(t)\right] dt, \tag{18}$$

where $\bar{\Phi}_1^{-1}$ (resp. $\bar{\Phi}_2^{-1}$) is the QF associated with the Fréchet mean distribution based on the ℓ_2 Wasserstein metric $M_W(Y_1)$ (resp. $M_W(Y_2)$).

For the i-th and the j-th generic units, we modify the indices of $\rho(\cdot,\cdot)$ in Eq. (16) such that $\rho_{i1,j2}$ denotes the correlation of the QFs Φ_{i1}^{-1} and Φ_{j2}^{-1}, while $\rho_{\cdot 1,\cdot 2}$ indicates the correlation of the QFs associated with $M_W(Y_1)$ (i.e., $\bar{\Phi}_{i1}^{-1}$) and $M_W(Y_2)$ (i.e., $\bar{\Phi}_{j2}^{-1}$). Utilizing the notation and the product of two QFs as defined in Eq.(17), the empirical covariance $C_W(Y_1, Y_2)$ is expressed as follows:

$$\begin{aligned} C_W(Y_1, Y_2) &= \left(\frac{1}{n}\sum_{i=1}^{n} \mu_{i1}\mu_{i2} - \mu_{\bar{y}_1}\mu_{\bar{y}_2}\right) \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^{n} \rho_{i1,i2}\sigma_{i1}\sigma_{i2} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{i1,j2}\sigma_{i1}\sigma_{j2}\right). \end{aligned} \tag{19}$$

This formulation provides an innovative method for measuring the covariance and correlation between two symbolic variables, addressing some limitations of the previous approach by Billard and Diday (2007). The use of the ℓ_2 Wasserstein metric and quantile functions in this context allows for a more nuanced and accurate analysis of symbolic data.

4.3 An example

We use the functions `hist_mean`, `hist_var`, `hist_cov`, and `hist_cor` to illustrate the process of obtaining basic descriptive statistics for the histogram-valued variables `Cholesterol` and `Hemoglobin` from the `BLOOD` dataset, using the BG (Bertrand and Goupil), BD (Billard and Diday), B (Billard) and L2W (ℓ_2 Wasserstein) approaches. The detailed results are presented in Table 8. The basic descriptive statistics computed using four different approaches show consistent results for the mean of `Cholesterol` across all methods (180.6770). However, notable differences are observed in the variance, covariance, and correlation estimates. The variance values range from 374.8639 (BG) to 705.1574 (BD), indicating that the BD approach yields a substantially higher variance compared to the others. Covariance estimates are negative across all methods, with the BD approach producing the most negative value (-8.5635). In terms of correlation, while the BG and BD approaches yield negative correlations (-0.2213 and -0.3659, respectively), the B and L2W approaches result in positive correlations (0.2005 and 0.4795, respectively), suggesting methodological differences in capturing the association between `Cholesterol` and `Hemoglobin`.

5 Pros and cons

The **dataSDA** package aims to provide comprehensive datasets and basic statistical tools for Symbolic Data Analysis (SDA) in R, addressing the growing need for specialized resources to store and handle complex, aggregated data structures. Below, we discuss the key strengths and limitations of the package.

5.1 Strengths

The **dataSDA** package offers several key strengths, making it a valuable tool for symbolic data analysis (SDA). First, it provides a comprehensive dataset collection, featuring a wide range of curated datasets spanning economics, social sciences, bioinformatics, and other domains, ensuring applicability to real-world problems with both classic and modern data. Second, its user-friendly interface includes intuitively named and well-documented functions, enabling users to efficiently load datasets, convert format, conduct analyses, and visualize results, while seamless integration with packages like **ggESDA** and **RSDA** further enhances accessibility. Third, the package's flexibility and extensibility support various symbolic data types—such as intervals, histograms, and modal data—allowing analysis of complex structures beyond traditional statistical methods, with room for custom technique implementation. Additionally, **dataSDA** has significant educational value, offering clear examples and tutorials to help students and researchers grasp SDA's theoretical foundations and practical applications. Lastly, its smooth integration with the broader R ecosystem leverages existing tools for data manipulation, statistical analysis, and visualization, minimizing the learning curve for R users. Together, these strengths position **dataSDA** as a robust and versatile resource for SDA research and education.

5.2 Limitations

Despite its strengths, the **dataSDA** package has several limitations that warrant consideration. First, while we have incorporated diverse symbolic datasets, the collection may not equally represent all potential application domains, potentially requiring additional specialized datasets for certain fields. For example, domains such as multilingual speech synthesis, text-to-speech (TTS) systems with sentiment analysis, and battery life prediction offer promising applications for symbolic dataset generation [56]. Second, although the package supports fundamental symbolic data formatting and basic statistical operations, it lacks a unified framework for advanced analytical methods such as symbolic regression, clustering, and machine learning, which may limit its utility for sophisticated analyses. Third, while experienced R users will find the package intuitive, its accessibility could be improved through supplementary resources like video tutorials or a graphical user interface to accommodate less technical users.

Additionally, while documentation and examples are provided, they may prove insufficient for complex use cases, suggesting a need for expanded tutorials and practical case studies. The current version primarily focuses on descriptive statistics rather than incorporating more advanced techniques like predictive modeling or machine learning approaches for symbolic data. Finally, while leveraging external R packages enhances functionality, this dependency introduces potential compatibility risks with future R updates. Addressing these limitations would provide important context for users, broaden adoption, and help guide future development priorities.

6 A benchmarking study

This section conducts a benchmarking study, evaluating clustering, classification, and regression algorithms applied to symbolic datasets within the **dataSDA** package. All algorithms are utilized with their default parameters.

6.1 Cluster analysis

Cluster analysis, a technique widely employed in data analysis, aims to reveal inherent structures and patterns within a dataset by grouping similar data objects into distinct clusters or groups. This method fundamentally seeks to create clusters without the need for predefined categories or labels, instead leveraging the intrinsic similarity within the data. Throughout the cluster analysis process, similarity measures play a crucial role, quantifying the resemblance between two data objects, typically computed using distance metrics or similarity indices. In this section, we utilize datasets from the **dataSDA** package to benchmark the clustering algorithm designed for interval-valued and histogram-valued datasets.

Cluster analysis for interval-valued data The **RSDA** package provided `sym.kmeans` function that extend the K-means algorithm to interval data. The quality of the resulting clusters is evaluated by $1 - WSS/TSS$, where WSS is the within sum of squares and

TSS is the total sum of squares (the higher the better). Lauro et al. [45] introduces two clustering algorithms for interval-valued data: Symbolic Clustering Algorithm (SClust) and Symbolic Clustering Algorithm on Distance Tables (DClust), which have been implemented in **symbolicDA** package. The primary difference between them lies in the input data. SClust utilizes symbolic objects (SOs) to model concepts, while DClust takes a distance matrix as input. In SClust, the assignment of elements to classes depends on the nature of the variables describing SOs, specifically tailored to the chosen prototype type for class representation. In DClust, distances between SOs are computed based on their descriptions, and class assignment relies on these distances. Apart from input data and assignment methods, SClust is typically employed when the description space for concepts and prototypes is homogenous, meaning that both SOs and prototypes share the same variable types, such as all being categorical or interval variables. In contrast, DClust is preferred when the description spaces for concepts and prototypes are not homogenous. This occurs when concepts are described by categorical, multi-valued, or interval variables, while prototypes are described using modal variables.

Table 9 reports the quality indices for K-means, DClust, and SClust applied to interval-valued datasets. The optimal number of clusters is determined by the majority vote across the three algorithms. Overall, `sym.kmeans` demonstrates superior performance across the majority of datasets, achieving the highest quality index on 22 out of the 28 datasets evaluated. In contrast, `SClust` shows competitive results in a few cases, notably outperforming the other methods on the `bird` and `Environment` datasets, and matching performance on datasets such as `baseball`, `Cardiological`, `facedata`, and `oils`. `DClust`, while occasionally comparable, ranks lower in most cases, showing the best performance on only one dataset (`soccer.bivar`). The performance gap is particularly pronounced in datasets like `blood_pressure`, `Abalone`, and `VeterinaryData`, where `sym.kmeans` achieves substantially better results. These findings highlight `sym.kmeans` as the most robust and consistently effective clustering approach among the methods tested for symbolic interval-valued data.

Cluster analysis for histogram-valued data Cluster analysis of histogram-valued data presents several challenges, particularly when compared to traditional data. A primary challenge arises from the multi-valued representation of each descriptor in histogram-valued data, in contrast to conventional data, which is described by a single value for each variable. As a result, distance metrics used in the cluster analysis of histogram-valued data must accommodate the multiple values associated with each descriptor. Another challenge is rooted in the bounded nature and local uniform distribution of histogram-valued data, necessitating that distance metrics employed in its cluster analysis account for these characteristics. Traditional distance metrics, such as Euclidean distance or cosine similarity, may not be apt for histogram-valued data due to their inability to consider the data's boundedness and uniformity.

Irpino and Verde (2006) [38] introduced an innovative distance metric that utilizes the Wasserstein metric, designed for clustering histogram-valued symbolic data. The Wasserstein metric, which accounts for the presence of multiple values within each descriptor, forms the basis for an agglomerative hierarchical clustering method tailored for histogram

data. The **HistDAWass** package offers five clustering algorithms for histogram-valued data: K-means (`WH_kmeans`), K-means using adaptive Wasserstein distances (`WH_adaptive.kmeans`), Fuzzy c-means (`WH_fcmeans`), Fuzzy c-means using adaptive Wasserstein distances (`WH_adaptive_fcmeans`) and hierarchical clustering based on the L2 Wasserstein distance (`WH_hclust`). We use K-means and Fuzzy C-means as demonstrative examples because the **HistDAWass** package provides quality indices by default.

Table 10 displays the quality indices ($1 - WSS/TSS$) for these two clustering algorithms applied to the seven histogram-valued datasets. The results show that both algorithms perform similarly overall, with slight variations depending on the dataset. `WH_fcmeans` outperforms `WH_kmeans` on four datasets `Age_Pyramids_2014`, `BLOOD`, `BloodBRITO`, and `China_Month`, with particularly marginal improvements in `BLOOD` and `BloodBRITO`. On the other hand, `WH_kmeans` shows better results on the remaining three datasets (`Agronomique`, `China_Seas`, and `OzoneFull`), most notably on `Agronomique`, where the difference in quality index is more pronounced. Overall, both clustering methods provide competitive results, with no clear dominant algorithm across all cases, suggesting their performance may vary depending on the structure and characteristics of the histogram-valued data.

6.2 Classification for interval-valued data

Classification analysis seeks to construct models from labeled data, enabling the assignment of new, unlabeled data points to predefined categories or labels. The process includes data preprocessing, feature selection or extraction, model training, and evaluation. Model assessment is of paramount importance, typically utilizing metrics like accuracy, precision, recall, and F1 score to evaluate model performance. This study adopts accuracy for simplicity.

The **symbolicDA** package provides several classification algorithms for symbolic data, including the optimal split-based decision tree (`decisionTree.SDA`), the random forest algorithm for optimal split-based decision trees (`random.forest.SDA`), and bagging/boosting algorithms for optimal splits derived from decision trees (`bagging.SDA` and `boosting.SDA`). Additionally, the **MAINT.Data** package implements linear discriminant analysis (`lda`) and quadratic discriminant analysis (`qda`) for interval-valued data classification. Table 11 compares the prediction accuracy of three selected algorithms, `decisionTree.SDA`, `lda`, and `qda`, when applied to suitable interval-valued datasets from **dataSDA** for classification tasks.

Overall, the `lda` classifier demonstrates the highest and most consistent accuracy across the majority of datasets, achieving particularly strong performance on `Cars` and `ohtemp` (both at 0.7950 or above), and outperforming other methods in five out of seven cases. `qda` also performs well but generally trails slightly behind `lda`, while `decisionTree.SDA` shows noticeably lower accuracy on most datasets. The only exception is the `lackinfo` dataset, where `decisionTree.SDA` slightly surpasses `lda` and clearly outperforms `qda`. The results indicate that while symbolic-specific classifiers like `decisionTree.SDA` are useful, traditional statistical classifiers (`lda` and `qda`) applied to summary statistics can often provide superior predictive performance for symbolic data, particularly when class

distributions are more balanced and well-defined.

6.3 Regression analysis for interval-valued data

In regression analysis, the relationship between an interval-valued dependent variable and interval-valued independent variables is modeled using a regression equation. When dealing with symbolic data, it is imperative to substitute mean, variance, covariance, and correlation functions with their symbolic statistical equivalents. Furthermore, it is assumed that these intervals are uniformly distributed. This assumption enables regression line fits to align with the results derived from applying classical methods to the midpoints of the intervals [7].

The **RSDA** package provides multiple methods for regression analysis of interval-valued data, including linear regression using the center method (CR) and center and range method (CRM) (`sym.lm`), symbolic bivariate regression (`sym.lm.bi`), regularized linear regression (lasso, ridge, and elastic net) (`sym.glm`), symbolic k-nearest neighbor regression (`sym.knn`), symbolic neural network regression (`sym.nnet`), symbolic random forest regression (`sym.rf`), symbolic regression trees (`sym.rt`), and symbolic support vector machines regression (`sym.svm`). Table 12 presents the coefficients of determination for two selected regression methods, evaluated on interval-valued datasets from **dataSDA** that are suitable for regression analysis.

The results show that `sym.lm` generally outperforms `sym.rf` in terms of R^2, with higher predictive accuracy on most datasets. In particular, `sym.lm` achieves excellent fit ($R^2 > 0.9$) on datasets such as `Abalone`, `ChinaTemp`, `nycflights`, `Environment`, `horses`, and `bird`, indicating strong linear relationships in those cases. However, `sym.rf` surpasses `sym.lm` on a few datasets, such as `Cardiological2`, `facedata`, and `baseball`, suggesting that random forests may better capture nonlinear or complex patterns in certain contexts. Nevertheless, in the majority of cases, including datasets with multiple predictors, `sym.lm` shows greater reliability and higher R^2, underscoring its effectiveness for interval-valued regression tasks within symbolic data analysis.

7 Conclusion and future developments

This study introduces the development of an R package, **dataSDA**, primarily designed to facilitate the collection of symbolic datasets, enable format conversion among different symbolic classes, and implement various approaches to calculate descriptive statistics for symbolic data. Benchmark analyses for classification, clustering, and regression have been performed using various datasets within **dataSDA**. Concurrently, benchmark analysis on the histogram-valued data in **dataSDA** has been specifically conducted for clustering analysis. This specificity arises due to the limited classification and regression methods proposed for histogram-valued data, and the rarity of such methods being implemented in R.

In comparison to previous reference works in symbolic data analysis such as Bock and Diday (2000) [10] and Billard and Diday (2003) [8] which primarily focused on theoretical foundations and small illustrative datasets, the **dataSDA** package provides a modern,

practical resource that is tightly integrated with the R environment. Unlike earlier efforts, which often lacked standardized formats and accessibility, **dataSDA** emphasizes reproducibility, standardized structure, and ease of use, thereby filling a crucial gap between symbolic data theory and applied data analysis workflows.

7.1 Key contributions

The proposed **dataSDA** package offers several key contributions and advantages over existing symbolic data resources. First, it compiles a diverse and growing collection of symbolic datasets, both interval- and histogram-valued, accompanied by consistent metadata and format compatibility with existing SDA packages. Second, it includes tools for computing symbolic descriptive statistics, enabling users to explore and summarize symbolic data in a structured and automated way. Third, by providing datasets commonly used in published SDA studies, the package facilitates benchmarking, comparison, and validation of new algorithms, thus enhancing scientific rigor and transparency. Finally, the integration within the R ecosystem ensures wide accessibility for the data science and statistics communities.

The provision of open-access symbolic datasets enhances reproducibility in research, enabling researchers to validate and build upon existing work. **dataSDA** aims to serve as an invaluable resource for students and professionals exploring machine learning, data analysis, and related fields, by offering real-world datasets for practical learning and application, especially for R users. Researchers who contribute to and utilize this database can amplify the visibility and impact of their work within the scientific community. Furthermore, the **dataSDA** package supplies datasets that are commonly employed for benchmarking new SDA algorithms and methods, thereby lending credibility to studies that utilize its datasets.

7.2 Future developments

We plan to significantly expand the capabilities of the **dataSDA** package in several directions. First, we aim to incorporate advanced symbolic data analysis techniques, including additional distance and similarity metrics (e.g., symbolic Wasserstein distance, histogram matching measures) to enable more sophisticated pattern recognition and machine learning applications. We will also implement descriptive symbolic statistics based on likelihood and Bayesian approaches, reflecting recent developments in the field.

Regarding data accessibility and community engagement, we intend to develop a web-based platform to complement the R package. This platform will serve multiple purposes: (i) providing plain-text access to datasets for non-R users, (ii) facilitating dataset donation by researchers through a standardized submission system, and (iii) promoting discussions and collaborations among the symbolic data analysis community. To ensure consistent growth, the package maintainer will regularly incorporate new, high-quality datasets, verify data integrity, and provide detailed metadata and documentation for each dataset.

We will systematically gather and standardize symbolic datasets from diverse domains (e.g., biomedical, environmental, and social sciences) through collaborations with domain experts. To better support symbolic data analysis education, we plan to expand our col-

lection of teaching materials, including case studies, video tutorials, and classroom-ready exercises. Additionally, we aim to introduce features for generating simulated interval-valued and histogram-valued datasets to facilitate method evaluation and benchmarking. Finally, future versions of **dataSDA** will incorporate evaluation metrics to support the comparative analysis of new symbolic data analysis methods and algorithms. These developments will further enhance **dataSDA**'s utility as a comprehensive research tool and educational resource while maintaining its core strengths in data accessibility, quality, and symbolic data analysis support.

Acknowledgments

The authors extend their heartfelt thanks to the anonymous researchers for their generous donation of the data sets. To donate symbolic datasets, researchers can submit their data via the website at `https://hmwu.idv.tw/dataSDA/`. This research received support from the National Science and Technology Council, Taiwan, R.O.C. (NSTC 113-2118-M-004-001, NSTC 113-2118-M-001-009-MY2). The **dataSDA** package is currently available on the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=dataSDA`.

Declaration of interest statement

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Disclosure of AI use

ChatGPT5 (OpenAI) was used solely for language editing (sentence polishing and proofreading). All content was reviewed and verified by the authors, who take full responsibility for the work.

References

[1] Bean B. Intkrige: a numerical implementation of interval-valued kriging. R package version 1.0.1;2020.

[2] Bean B, Maguire M, Sun Y. The Utah snow load study. Civil and Environmental Engineering Faculty Publications. 2018; Paper 3589.

[3] Beranger B, Lin H, & Sisson S. New models for symbolic data analysis. Adv Data Anal Classif. 2023;17(3):659–699.

[4] Bertrand P, Goupil F. Descriptive statistics for symbolic data. In: Analysis of Symbolic Data, Bock HH, Diday E. (eds). Springer, Berlin, Heidelberg. 2000;106–124.

[5] Billard L. Dependencies and variation components of symbolic interval-valued data. In: Selected contributions in data analysis and classification. Springer. 2007;3–12.

[6] Billard L. Sample covariance functions for complex quantitative data. In: Proceedings of World IASC Conference, Yokohama, Japan. 2008;157–163.

[7] Billard L, Diday E. Regression analysis for interval-valued data. In: Data Analysis, Classification, and Related Methods. Springer. 2000;369–374.

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

[9] Billard L, Diday E. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Ltd; 2007.

[10] Bock HH, Diday E. Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Springer. 2000.

[11] Borcard D, Gillet F, Legendre P. Numerical Ecology with R. Springer New York; 2011.

[12] Borysov SS, Geilhufe RM, Balatsky AV. Organic materials database: An open-access online database for data mining. PLoS ONE. 2017;12(2): e0171501.

[13] Brito P, Duarte Silva AP. Modelling interval data with normal and skew-normal distributions. J. Appl. Stat. 2012;39(1):3–20.

[14] Cazes P, Chouakria A, Diday E, Schechtman Y. Extension de l'analyse en composantes principales 'a des donn'ees de type intervalle. Rev Stat Appl. 1997;45, 5–24.

[15] Chiang K, Shu J, Zempleni J, Cui J. Dietary microRNA database (DMD): an archive database and analytic tool for food-borne microRNAs. PLoS ONE. 2015;10(6):e0128089.

[16] Chouakria A. Extension de l'analyse en composantes principales 'a des donn'ees de type intervalle." Doctoral Thesis;University of Paris IX Dauphine; 1998.

[17] Chouakria A, Cazes P, Diday E. Symbolic principal component analysis," In: Analysis of Symbolic Data, Bock HH, Diday E (eds). Berlin, Springer-Verlag; 2000.

[18] Dau HA, Keogh E, et al. The UCR time series classification archive. 2019. URL https://www.cs.ucr.edu/ eamonn/time_series_data_2018/

[19] De Carvalho FdA. Fuzzy c-means clustering methods for symbolic interval data. Pattern Recognit. Lett. 2007;28(4):423–437.

[20] DeCarvalho, FdA, Lechevallier Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. Pattern Recognit. 2009;42(7):1223–1236.

[21] Denoeux T, Masson M. Multidimensional scaling of interval-valued dissimilarity data. Pattern Recognit. Lett. 2000;21(1):83–92.

[22] Douzal-Chouakria A, Billard L, Diday E. Principal component analysis for interval-valued observations. Stat Anal Data Min. 2011;4(2):229–246.

[23] Diday E. The symbolic approach in clustering and related methods of data analysis: the basic choices. In: Classification and Related Methods of Data Analysis, Proceedings of the First Conference of the International Federation of Classification Societies. IFCS-87: Technical University of Aachen. North Holland. 1988;673–684.

[24] Diday E, Noirhomme-Fraiture M. Symbolic Data Analysis and the SODAS Software, Wiley-Interscience.; 2008.

[25] Diday, E. (2016). Thinking by classes in data science: the symbolic data analysis paradigm. Wiley Interdisciplinary Reviews: Computational Statistics;8(5):172–205.

[26] D'Urso P, Giordani P. A least squares approach to principal component analysis for interval valued data. Chem Intell Lab Syst. 2004;70:179–192.

[27] Kelly M, Longjohn R, Nottingham K, The UCI Machine Learning Repository. 2023; https://archive.ics.uci.edu

[28] Garcia J. IntervalQuestionStat: tools to deal with interval-valued responses in questionnaires. R package version 0.1.0; 2022.

[29] Gilchrist W. Statistical Modelling with Quantile Functions. Chapman & Hall; 2000.

[30] Gioia F, Lauro NC, Principal component analysis on interval data. Comput. Stat. 2006;21:343–363.

[31] Groenen PJF, Winsberg S, Rodriguez O, Diday E. I-Scal: multidimensional scaling of interval dissimilarities. Comput Stat Data Anal. 2006;51(1):360–378.

[32] Grzegorzewski P, Śpiewak M. The sign test and the signed-rank test for interval-valued data. Int. J. Intell. Syst. 2019;34(9):2122–2150.

[33] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. Springer (2nd edition); 2009.

[34] Hayes B, A lucid interval. Am. Sci. 2003;91(6):484–488.

[35] Henderson HV, Velleman PF. Building multiple regression models interactively. Biometrics. 1981;37(2):391–411.

[36] Ichino M. The quantile method for symbolic principal component analysis. Stat Anal Data Min. 2011;4(2):184–198.

[37] Irpino A. "Spaghetti" PCA analysis: an extension of principal components analysis to time dependent interval data. Pattern Recognit. Lett. 2006;27:504–513.

[38] Irpino A, Verde R. A new Wasserstein-based distance for the hierarchical clustering of histogram symbolic data. In: Data Science and Classification. Studies in Classification, Data Analysis, and Knowledge Organization, Batagelj V, Bock HH, Ferligoj A, Žiberna A. (eds). Springer, Berlin, Heidelberg. 2006;185–192.

[39] Irpino A, Verde R. Basic statistics for distributional symbolic variables: a new metric-based approach. Adv Data Anal Classif. 2015;9:143–175.

[40] Irpino A, Verde R, De Carvalho FdA. Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications. 2014;41(7):3351–3366.

[41] Kao CH, Nakano J, Shieh SH, Tien YJ, Wu HM, Yang CK, Chen CH. Exploratory data analysis of interval-valued symbolic data with matrix visualization. Comput Stat Data Anal. 2014;79:14–29.

[42] Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GPS. TumorHoPe: A database of tumor homing peptides. PLoS ONE. 2012;7(4):e35187.

[43] Lauro CN, Palumbo F. Principal component analysis of interval data: a symbolic analysis approach. Comput. Stat. 2000;15(1):73–87.

[44] Lauro CN, Gioia F. Dependence and interdependence analysis for interval-valued variables. In: Data Science and Classification, Batagelj V, HBock HH, Ferligoj A, Ziberna A (eds). Berlin, Springer-Verlag. 2006;171–183.

[45] Lauro NC, Verde R, Irpino A. Principal component analysis of symbolic data described by intervals. In: Symbolic Data Analysis and the SODAS Software, Diday E, Noirhomme-Fraiture M (eds). Wiley, Chichester. 2008;279–311.

[46] Lauro NC, Verde R, Palumbo F. Factorial data analysis on symbolic objects under cohesion constrains. In: Data Analysis, Classification and Related Methods. Springer-Verlag, Heidelberg; 2000.

[47] Lee JA, Verleysen M. Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing. 2009;72:1431–1443.

[48] Lee JA, Verleysen M. Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. JMLR: Workshop and Conference Proceedings. 2008;4: 21–35.

[49] Lee JA, Verleysen M. Scale-independent quality criteria for dimensionality reduction. Pattern Recognit. Lett. 2010;31:2248–2257.

[50] Leroy B, Chouakria A, Herlin I, Diday E. Approche geometrique et classification pour la reconnaissance de visage, Reconnaissance des Forms et Intelligence Artificelle, INRIA and IRISA and CNRS, France. 1996;548–557.

[51] Le-Rademacher J, Billard L. Symbolic covariance principal component analysis and visualization for interval-valued data. J Comput Graph Stat. 2012;21(2):413 -– 432.

[52] Meng D, Leung Y, Xu Z. A new quality assessment criterion for nonlinear dimensionality reduction. Neurocomputing. 2011;74:941–948.

[53] Mokbel B, Lueks W, Gisbrecht A, Hammer B. Visualizing the quality of dimensionality reduction. Neurocomputing. 2013;112:109–123.

[54] Neto EAL, Cordeiro GM, de Carvalho FdA. Bivariate symbolic regression models for interval - valued variables. J Stat Comput Simul. 2011;81(11):1727–1744.

[55] Neto EAL, de Carvalho FdA. Centre and range method for fitting a linear regression model to symbolic interval data. Comput Stat Data Anal. 2008;52(3):1500–1515.

[56] Nuthakki P, Katamaneni M, JN CS, Gubbala K, Domathoti B, Maddumala VR, & Jetti KR. Deep learning based multilingual speech synthesis using multi feature fusion methods. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 2023.

[57] Palumbo F, Lauro CN. A PCA for interval valued data based on midpoints and radii, In: New Developments in Psychometrics, Yanai H, Okada A, Shigematu K, Kano Y, Meulman JJ (eds). Japan, Springer-Verlag. 2003;641–648.

[58] Rüschendorf L. Wasserstein metric. In: Encyclopaedia of Mathematics, Hazewinkel M (ed), Springer; 2001.

[59] Silva APD, Brito P. Discriminant analysis of interval data: An assessment of parametric and distance-based approaches. J. Classif. 2015;32:516–541.

[60] Silva APD, Brito P, Filzmoser P, Dias JG. MAINT.Data: modelling and analysing interval data in R. The R Journal. 2021;13(2):336–364.

[61] Su ECY, & Wu HM. Dimension reduction and visualization of multiple time series data: a symbolic data analysis approach. Comput. Stat. 2024;39(4):1937–1969.

[62] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290:2319–2323.

[63] Umbleja K, Ichino M, Yaguchi H. Improving symbolic data visualization for pattern recognition and knowledge discovery. Visual Informatics. 2020;4(1):23–31.

[64] Verde R, Batagelj V, Brito P, Silva APD, Korenjak-Černe S, Dobša J, & Diday E New skills in symbolic data analysis for official statistics. Statistical Journal of the IAOS. 2024;40(3):563–579.

[65] Verde R, Irpino A. Dynamic clustering of histogram data: using the right metric. In: Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization, Brito P, Cucumel G, Bertrand P, de Carvalho F. (eds). Springer, Berlin, Heidelberg. 2007;123–134.

[66] Wang H, Guan R, Wu J. CIPCA: Complete-information-based Principal Component Analysis for interval-valued data. Neurocomputing. 2012;86:158–169.

[67] Wickham et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686.

[68] Xu W. Symbolic Data Analysis: Interval-valued Data Regression. PhD thesis, University of Georgia Athens, GA; 2010.

[69] Zhang P, Ren Y, Zhang B. A new embedding quality assessment method for manifold learning. Neurocomputing. 2012;97:251–266.

Table 2: Summaries of the interval-valued data sets in **dataSDA** package.

Dataset	**Subject Area**	**Tasks**	**Size**	**Reference**
`Abalone`	Life Science	Regression	24×7	[4]
`age_cholesterol_weight`	Life Science	Regression	7×4	[5]
`baseball`	Life Science	Classification	19×3	[5]
`bird`	Life Science	Other	20×2	[5]
`blood_pressure`	Life Science	Regression	15×3	[5]
`Cardiological`	Life Science	Regression	11×3	[5, 68]
`Cardiological2`	Life Science	Regression	15×3	[5, 68]
`Cars`	Other	Classification	27×5	[5]
`ChinaTemp`	Life Science	Classification	899×5	[60]
`Environment`	Life Science	Regression	14×17	[63]
`facedata`	Life Science	Regression	27×6	[22, 51]
`finance`	Business	Regression	14×7	[5]
`horses`	Life Science	Regression	8×7	[5]
`iris.i`	Life Science	Regression	3×4	[5]
`lackinfo`	Life Science	Classification	50×8	[28]
`LoansbyPurpose`	Life Science	Regression	14×4	`kaggle`
`LoansbyRiskLvs`	Life Science	Regression	35×4	`kaggle`
`lynne1`	Life Science	Regression	10×4	[5]
`mtcars.i`	Other	Regression	5×11	[35]
`mushroom`	Life Science	Classification	23×5	[5]
`nycflights`	Other	Regression	142×5	[60]
`ohtemp`	Life Science	Classification	161×7	[1]
`oils`	Other	Other	8×4	[5]
`profession`	Social Science	Classification	15×4	[5]
`soccer.bivar`	Other	Regression	20×3	[54]
`Uscrime`	Social Science	Regression	46×102	[33]
`utsnow`	Life Science	Other	415×7	[2]
`VeterinaryData`	Life Science	Other	10×2	[5]

Table 3: Summaries of the histogram-valued data sets in **dataSDA** package.

Data	**Subject Area**	**Tasks**	**Size**	**Reference**
`Age_Pyramids_2014`	Life Science	Regression	229×3	U.S. Census Bureau
`Agronomique`	Life Science	Regression	22×4	[40]
`airline_flights2`	Other	Other	16×6	[5]
`BLOOD`	Life Science	Regression, Clustering	14×3	[5]
`BloodBRITO`	Life Science	Other	10×2	[5]
`China_Month`	Life Science	Regression, Clustering	60×168	[40]
`China_Seas`	Life Science	Regression, Clustering	60×56	[40]
`OzoneFull`	Life Science	Regression, Clustering	78×4	[40]
`OzoneH`	Life Science	Regression, Clustering	84×4	[40]

Table 4: The descriptive statistic provided by the various R packages available on CRAN for interval-valued data.

Package	Mean	Variance	Covariance	Correlation
dataSDA	`int_mean`	`int_var`	`int_cov`	`int_cor`
ggESDA	–	–	`cor`	`cov`
IntervalQuestionStat	`mean`	`var`	`cov`	–
psda	`pmean_id`	`pvari`	`pcov`	`pcorr`
RSDA	`mean`	`var`	`cov`	`cor`

Table 5: The mean and variance of the variable `Pileus.Cap.Widthh`, as well as the covariance and correlation between the variables `Pileus.Cap.Width` and `Stipe.Length` within the `mushroom` dataset based on the various approaches.

	Quantification					Distributional		
Statistics	CM	VM	QM	SE	FV	EJD	GQ	SPT
Mean	7.9783	7.9783	7.9783	7.9783	11.1957	7.9783	7.9783	7.9783
Variance	11.4654	26.0329	18.3768	26.0329	13.8127	15.8003	15.8003	15.8003
Covariance	8.4180	8.1405	14.0218	20.1865	10.3720	8.0520	11.4609	11.9505
Correlation	0.8047	0.3556	0.8578	0.8818	0.7497	0.5695	0.8106	0.8453

Table 6: The mean and variance of the variable `Sepal.Length`, as well as the covariance and correlation between the variables `Sepal.Length` and `Sepal.Width` within the `iris` dataset based on the various approaches.

	Quantification					Distributional		
Statistics	CM	VM	QM	SE	FV	EJD	GQ	SPT
Mean	5.8433	5.8433	5.8433	5.8433	5.8433	5.8433	5.8433	5.8433
Variance	0.6857	0.6857	0.6819	0.6834	0.6857	0.6811	0.6811	0.6811
Covariance	-0.0424	-0.0424	-0.0422	-0.0423	-0.0424	-0.0422	-0.0422	-0.0422
Correlation	-0.1176	-0.1176	-0.1176	-0.1176	-0.1176	-0.1176	-0.1176	-0.1176

Table 7: The functions for calculating descriptive statistics for histogram-valued variables are provided by several R packages available on CRAN.

Packages	Mean	SD/Var.	Covariance	Correlation	Other statistics
dataSDA	`hist_mean`	`hist_var`	`hist_cov`	`hist_cor`	-
HistDat	`mean`	`sd`	`var`	-	`length`, `max`, `min`, `median`, `quantile`, `range`, `sum`
HistDAWass	`get.m`	`get.s`	`WH.var.covar`	`WH.correlation`	`get.MatH.stats`
	`meanH`	`stdH`	`WH.var.covar2`	`WH.correlation2`	

Table 8: The mean and variance of the variable `Cholesterol`, as well as the covariance and correlation between the variables `Cholesterol` and `Hemoglobin` within the `BLOOD` dataset based on the various approaches.

Statistics	BG	BD	B	L2W
mean	180.6770	180.6770	180.6770	180.6770
var	374.8639	705.1574	400.0263	388.1376
cov	-5.1790	-8.5635	-4.6927	-5.0005
cor	-0.2213	-0.3659	0.2005	0.4795

Table 9: The quality indices, defined as $1 - WSS/TSS$, of the clustering algorithms for interval-valued data are evaluated using datasets from the **dataSDA** package.

	Optimal	**RSDA**	**symbolicDA**	
Datasets	cluster number	`sym.kmeans`	`DClust`	`SClust`
`Abalone`	3	0.8430	0.5225	0.6612
`age_cholesterol_weight`	2	0.8180	0.6352	0.7842
`baseball`	2	0.6821	0.6770	0.6770
`bird`	4	0.8760	0.8726	0.9197
`blood_pressure`	4	0.9150	0.6473	0.5731
`Cardiological`	3	0.8670	0.7233	0.7233
`Cardiological2`	4	0.7100	0.6508	0.5788
`Cars`	3	0.8970	0.8403	0.8471
`ChinaTemp`	4	0.8760	0.8570	0.8432
`Environment`	5	0.7450	0.5371	0.8471
`facedata`	4	0.7430	0.6867	0.6867
`finance`	3	0.9480	0.8712	0.7539
`horses`	3	0.8180	0.7588	0.6081
`iris.i`	2	0.8580	0.7762	0.7413
`lackinfo`	4	0.6250	0.6130	0.5864
`LoansbyPurpose`	3	0.7910	0.7058	0.7074
`LoansbyRiskLvs`	4	0.7670	0.7149	0.7340
`lynne1`	3	0.8160	0.7297	0.7995
`mtcars.i`	3	0.8520	0.8481	0.8517
`mushroom`	4	0.8610	0.7115	0.7330
`nycflights`	3	0.9430	0.8276	0.7709
`ohtemp`	3	0.8050	0.7462	0.7163
`oils`	3	0.9240	0.7991	0.7991
`profession`	3	0.9200	0.9046	0.9196
`soccer.bivar`	4	0.7100	0.7803	0.6314
`Uscrime`	5	0.5080	0.5172	0.5329
`utsnow`	3	0.8940	0.8462	0.8574
`VeterinaryData`	3	0.9830	0.9209	0.8242

Table 10: The quality indices, defined as $1 - WSS/TSS$, of three clustering algorithms for histogram-valued data are evaluated using datasets from the **dataSDA** package.

	Optimal	**HistDAWass**	
Datasets	cluster number	`WH_kmeans`	`WH_fcmeans`
`Age_Pyramids_2014`	2	0.7338	0.7489
`Agronomique`	3	0.6628	0.5385
`BLOOD`	2	0.6821	0.6824
`BloodBRITO`	3	0.8393	0.8406
`China_Month`	3	0.6236	0.5828
`China_Seas`	4	0.5667	0.5621
`OzoneFull`	4	0.5675	0.5424

Table 11: Accuracy of the three classifiers to the selected **dataSDA** datasets.

Datasets	Response: categories (no.)	**symbolicDA**	**MAINT.Data**	
		`decisionTree.SDA`	`lda`	`qda`
`baseball`	Pattern: A(1), B(9), C(1), D(1), E(1), F(1), G(1), H(1), I(1)	0.5263	0.7895	0.7368
`Cars`	class: Utilitarian(7), Berlina(8), Sportive(8), Luxury(4)	0.2963	0.9259	0.9259
`ChinaTemp`	GeoReg: East(225), North(75), Northeast(135), Northwest(194), South_central(120), Southwest(150)	0.2503	0.6952	0.5495
`lackinfo`	sex: female(32), male(18)	0.6400	0.6200	0.4400
`mushroom`	Edibility: T(4), U(2), Y(17)	0.7391	0.8261	0.7826
`ohtemp`	STATE: 1(10), 2(33), 3(18), 4(3), 5(2), 6(33), 7(18), 8(13), 9(5), 10(26)	0.2236	0.7950	0.7950
`profession`	Type_of_Work: Blue Collar(5), Services White(5), Collar(5)	0.3158	0.6000	0.6000

Table 12: The coefficient of determination (R^2) of regression models for selected interval-valued datasets in **dataSDA** package.

			RSDA	
Datasets	Response	Number of X's	`sym.lm`	`sym.rf`
`Abalone`	LENGTH	6	0.9893	0.9004
`age_cholesterol_weight`	Age	2	0.9588	0.5851
`baseball`	Hits	1	0.0443	0.4415
`bird`	Size	1	0.9294	0.9000
`blood_pressure`	Systolic_Pressure	2	0.4247	0.3281
`Cardiological`	Pulse	2	0.5777	0.5355
`Cardiological2`	Pulse	4	0.3772	0.8203
`Cars`	Price	3	0.9118	0.8176
`ChinaTemp`	Q1	3	0.9641	0.9637
`Environment`	INDIVIDUAL	12	0.9805	0.4168
`facedata`	AD	5	0.6379	0.8181
`finance`	Annual_Budget	4	0.5428	0.3807
`horses`	Minimum_Weight	5	0.9989	0.4559
`lackinfo`	item4	4	0.6327	0.3161
`LoansbyPurpose`	total-acc	3	0.8257	0.5784
`LoansbyRiskLvs`	ln-inc	3	0.4593	0.4278
`lynne1`	Systolic	2	0.7154	0.3486
`mushroom`	Pileus.Cap.Width	2	0.7437	0.6794
`nycflights`	distance	3	0.9912	0.9152
`oils`	GRA	3	0.9143	0.7379
`profession`	Salary	1	0.8184	0.7832
`VeterinaryData`	Weight	1	0.7348	0.6472