Programme and Abstracts

10TH WORKSHOP ON SYMBOLIC DATA ANALYSIS

SDA 2025

9 - 11 June 2025

University of Zagreb
Faculty of Organization and Informatics
Varaždin, Croatia

J. Dobša, M. Buhin Pandur
(Editors)

10th Workshop on Symbolic Data Analysis SDA 2025

Programme and Abstracts

9 – 11 June 2025
University of Zagreb Faculty of Organization and Informatics
Varaždin, CROATIA

Title:	*10th Workshop on Symbolic Data Analysis SDA 2025: Programme and Abstracts*
Publisher:	University of Zagreb Faculty of Organization and Informatics, Pavlinska Street 2, 42000 Varaždin, Croatia
For Publisher:	Prof. dr. sc. Marina Klačmer Čalopa
Editors:	Jasminka Dobša, University of Zagreb Faculty of Organization and Informatics Maja Buhin Pandur, University of Zagreb Faculty of Organization and Informatics
Cover Design:	Maja Buhin Pandur
ISBN:	978-953-6071-84-5
Number of Copies:	40

Preface

The 10th Workshop on Symbolic Data Analysis (SDA 2025) is being held in Varaždin, a city in northern Croatia, from June 9th to June 11th 2025, and is co-organized by the University of Zagreb Faculty of Organization and Informatics, and the Croatian Mathematical Society.

The goal of the Workshop is to provide a platform for networking among researchers working in the field of SDA, and to foster the exchange of ideas and recent advances in the area. On the first day of the event, prior to the official opening, tutorials and software demonstrations related to SDA are held. These sessions are organized in a hybrid format—both online and in person—with the aim of popularizing the field among young researchers, professionals from industry, and other interested participants.

We are witnessing the rapid development of the field of Data Science. The growing availability of data and the increasing computational power to process vast amounts of information in real time have opened unprecedented opportunities for gaining insights. At the same time, they have raised concerns about biased interpretations and the unexplanability of increasingly complex models. To develop data analysis models that are robust to these challenges, collaboration between the scientific community, industry, and public institutions—both national and international—is essential. As part of the effort to strengthen the collaboration between the Croatian Bureau of Statistics and the research community, a panel discussion titled "*New Horizons in Official Statistics: Techniques, Tools, and Challenges*" is included in the program.

The scientific program of the Workshop includes fundamental research and modeling in the field of SDA, investigations in established domains of data analysis using SDA tools, and applications in areas such as portfolio management and sentiment analysis. The chapters of this book of abstracts reflect the thematic sections of the Workshop: *Foundations of SDA and Statistical Modeling I*, *Dimensionality Reduction*, *Foundations of SDA and Statistical Modeling II*, *Clustering*, *Regression*, *Supervised Learning*, and *Applications of SDA*.

We would like to express our sincere gratitude to all authors for their contributions, and to the reviews for their valuable feedback and suggestions, which have significantly contributed to the quality of this Book of Abstracts.

Special thanks go to the University of Zagreb Faculty of Organization and Informatics, and the Croatian Mathematical Society for their generous support in organizing the event. We also gratefully acknowledge the sponsors, whose support helped make the participants' stay in Croatia more pleasant and welcoming.

Finally, we thank all participants for their interest in SDA. The continued development of this field depends on the engagement of a vibrant and curious community.

Varaždin, June 2025

Paula Brito
Simona Korenjak-Černe
Jasminka Dobša
Maja Buhin Pandur
Diana Šimić

Committees

Scientific Committee

Javier Arroyo, Complutense University of Madrid, Spain

Vladimir Batagelj, IMFM, Ljubljana, Slovenia and University Primorska, Koper, Slovenia

Lynne Billard, University of Georgia, USA

Paula Brito, University of Porto, Portugal

Francisco de Assis Tenório de Carvalho, Federal University of Pernambuco, Recife, Brazil

Chun-houh Chen, Academia Sinica, Taiwan

Simona Korenjak-Černe, University of Ljubljana, Slovenia

Jasminka Dobša, University of Zagreb, Croatia

Richard Emilion, University of Orléans, France

Antonio Irpino, University of Campania "Luigi Vanvitelli", Caserta, Italy

Ndèye Niang, National Conservatory of Arts and Crafts, Paris, France

Monique Noirhomme-Fraiture, University of Namur, Belgium

Oldemar Rodríguez, University of Costa Rica, Costa Rica

Scott Sisson, University of New South Wales, Sydney, Australia

Rosanna Verde, University of Campania "Luigi Vanvitelli", Caserta, Italy

Huiwen Wang, Beihang University, Beijing, China

Organising Committee

Paula Brito, University of Porto, Portugal

Simona Korenjak-Černe, University of Ljubljana, Slovenia

Jasminka Dobša, University of Zagreb, Croatia

Maja Buhin Pandur, University of Zagreb, Croatia

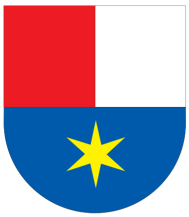
Diana Šimić, University of Zagreb, Croatia

Organisation

The Workshop is organised by the University of Zagreb Faculty of Organization and Informatics in collaboration with the Croatian Mathematical Society.

Support and Sponsors

With sincere appreciation, we acknowledge our sponsors whose generous support helped shape and realize SDA 2025.



MEĐIMURSKA ŽUPANIJA

Contents

Programme

Monday – June 9, 2025

08:30	**Registration**
09:15	**Participants Welcome**
09:30 – 11:00	**Tutorial (+ online)**
	Chair: Jasminka Dobša
	Symbolic Data Analysis – Why, How, What for?
	Paula Brito
11:00 – 11:30	**Coffee break**
11:30 – 13:00	**Software Presentations (+ online)**
	Chair: Renata M. C. R. de Souza
11:30 – 12:00	The RSDA Package
	Oldemar Rodríguez
12:00 – 12:30	R Package MAINT.Data
	Pedro Duarte Silva
12:30 – 13:00	HistDAWass: an R Package for the Exploratory Analysis of Histogram Data
	Antonio Irpino
13:00 – 14:15	**Lunch break**
14:15 – 14:30	**Workshop Opening**
14:30 – 16:30	**Session I: Foundations of SDA and Statistical Modelling I**
	Chair: Lynne Billard
14:30 – 15:00	File Representations of Symbolic Data for Open Science
	Vladimir Batagelj
15:00 – 15:30	Central Limit Theorem for a Set-valued Martingale
	Andrej Srakar
15:30 – 16:00	Multiblock Analysis of Distributional Data
	Paula Brito, Ndeye Niang, A. Pedro Duarte Silva, Stephanie Bougeard

16:00 – 16:30 Geometric Goodness-of-fit Measure for Interval-valued Data

Dylan Benavides, Oldemar Rodríguez, Maikol Solis

16:30 – 17:00 Coffee break

17:00 – 18:00 Session II: Dimensionality Reduction

Chair: M. Rosário Oliveira

17:00 – 17:30 Riemannian Principal Component Analysis for Interval-valued Data

Oldemar Rodríguez

17:30 – 18:00 Principal Component Analysis of Distributional Data

Sonia Dias, Paula Brito

18:15 – 20:00 Local Guided Tour of the City of Varaždin

20:00 – 22:00 Dinner at *Restaurant of the Park Boutique Hotel*

Tuesday – June 10, 2025

09:00 – 11:00 Session III: Foundations of SDA and Statistical Modelling II

Chair: Vladimir Batagelj

09:00 – 09:30 An Extension of Entropy for Interval-valued Data

Jose Andres, Piedra Molina, Oldemar Rodríguez Rojas

09:30 – 10:00 Advancements in the Best Point Method: New Optimization Criteria

Mario J. Gomez, Jorge Arce, Oldemar Rodríguez

10:00 – 10:30 Symbolic Data as Matrix-valued Data Robust Estimation and Explainable Outlier Detection

Marcus Mayrhofer, Paula Brito, A. Pedro Duarte Silva, Peter Filzmoser

10:30 – 11:00 Explainable Outlier Detection in Interval-valued Data Using a Robust Covariance Estimator

Catarina P. Loureiro, M. Rosário Oliveira, Paula Brito, Lina Oliveira

11:00 – 11:30 Coffee break

11:30 – 13:00 Session IV: Clustering

Chair: Simona Korenjak-Černe

11:30 – 12:00 Clustering Density-valued Data

Rui Nunes, Paula Brito, Sonia Dias

12:00 – 12:30 Clustering Intervals Using Principal Components

Jiankun Zhu, Lynne Billard

12:30 – 13:00 Clustering Distributional Data Using Mahalanobis-like Distance on LDQ Functions

Gianmarco Borrata, Rosanna Verde, Antonio Balzanella

13:00 – 14:30 Lunch break

14:30 – 15:30 Panel Discussion

Moderator: Jasminka Dobša

New Horizons in Official Statistics: Techniques, Tools, and Chalanges

Ivana Levačić, Jure Dubravčić, Boris Berenček, Paula Brito, Oldemar Rodríguez

16:00 – 22:00 Excursion and Dinner

Local Guided Tour Followed by Dinner in Međimurje County

Wednesday – June 11, 2025

09:00 – 10:30 Session V: Regression

Chair: Sonia Dias

09:00 – 09:30 From Ordinary Least Squares to Robust Methods: Revisiting Regression for Interval-valued Variables

M. Rosário Oliveira, Conceição Amado

09:30 – 10:00 Integrating Centre and Range Methods in Symbolic Regression Trees

Priscilla Angulo, Oldemar Rodríquez

10:00 – 10:30 Regression Models with Sentiment Analysis Integration for Interval-valued Data

Daniela Sevilla, Oldemar Rodríguez

10:30 – 11:00 Coffee break

11:00 – 12:00 Session VI: Supervised Learning

Chair: Pedro Duarte Silva

11:00 – 11:30 A New Discriminant Analysis Approach for Density-valued Data Classification

Francesca Condino, Paula Brito

11:30 – 12:00 Stacked Logistic Regression for Interval Data Classification

	Rafaella L. S. do Nascimento, Renata M. C. R. de Souza, Francisco José A. Cysneiros
12:00 – 13:00	**Session VII: Applications of SDA**
	Chair: Oldemar Rodríguez
12:00 – 12:30	LIMOS - LightGBM Interval Merton's One-period-portfolio Selection
	Liang-Ching Lin
12:30 – 13:00	Symbolic Data Analysis Approach to Identify Adolescent Profiles Based on Momentary Self-assessments and the Use of Internet Applications
	Jasminka Dobša, Simona Korenjak-Černe, Miranda Novak, Maja Buhin Pandur
13:00 – 13:30	**Closing**
	Summary, Announcements, Feedback
13:30 – 15:00	**Lunch**

Abstracts

Foundations of SDA and Statistical Modelling I

File Representations of Symbolic Data for Open Science

Vladimir Batagelj[1,2,3,*]

1. IMFM Ljubljana
2. UP FAMNIT Koper
3. UL FMF Ljubljana

*Contact author: vladimir.batagelj@fmf.uni-lj.si
ORCID: 0000-0002-0240-9446

Keywords: Open Science, FAIR, SDA file format, JSON, Standards.

In Open Science (Wikipedia, 2025), there is a growing emphasis on publishing research data following the FAIR principles (Findable, Accessible, Interoperable, Reusable) (GoFAIR, 2016). Adhering to these standards ensures the verifiability of the results and enables alternative analyses. Additionally, open data contributes to greater diversity in datasets, supporting the development and testing of new methodologies.

In symbolic data analysis, the starting point is usually a generalized (symbolic) data table, where variable values can be structured (combinations of primitive values). These require specialized external (file-based) and internal (in-memory) representations. Ideally, the two representations would be compatible.

This presentation focuses on file-based descriptions of symbolic data tables, which can facilitate seamless data exchange between symbolic data analysis tools.

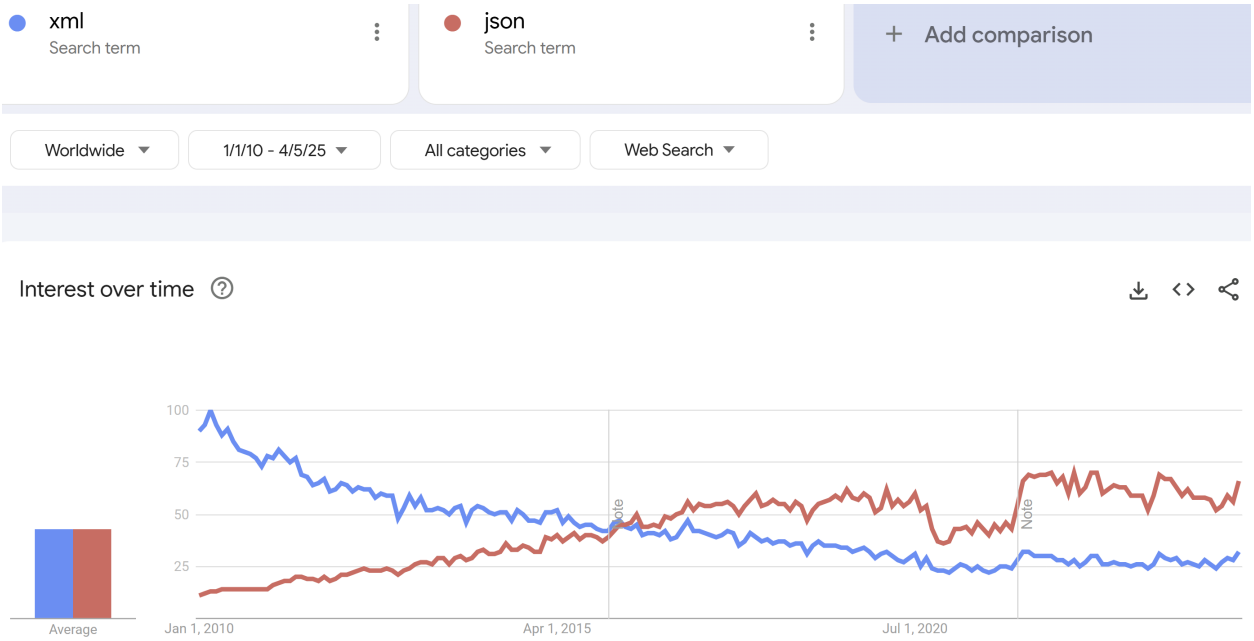


Figure 1: Google trends XML : JSON

Most formats for structured data are based on XML or JSON, with JSON increasingly favored in modern applications – see Figure 1. JSON description is not only a valid JavaScript expression but also uses data structures that are natively supported by most programming languages (e.g., *R*, *Python*, *Julia*, *C++*) (JSON, 2017; ECMAScript, 2024; Batagelj, 2016).

Beyond the raw data, it is essential to incorporate metadata in the file description. When designing such descriptions, it is advisable to rely on established standards, such as persistent identifiers (DOIs, ORCID, ROR) (DPC, 2025), ISO standards (ISO, 2025), schema.org (Schema.org, 2025), Dublin Core (DCMI, 2025), etc.

Adopting these practices ensures better interoperability, reusability, and long-term preservation of symbolic data.

The data and code will be available at GitHub/bavla symDATA/format.

References

Batagelj, V. (2016). Network visualization based on JSON and D3.js (slides). *Second European Conference on Social Networks. June 14-17, 2016, Paris*. https://github.com/bavla/netsJSON/blob/master/doc/netVis.pdf.

DCMI (2025). The Dublin Core Metadata Initiative. https://www.dublincore.org/.

Digital Preservation Coalition (2025). Persistent identifiers. In Digital Preservation Handbook. https://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers.

ECMA (2017). Introducing JSON – ECMA-404: The JSON data interchange syntax. 2nd edition. https://www.json.org/json-en.html.

ECMA (2024). ECMA-262: ECMAScript® 2024 Language Specification. 15th Edition. https://ecma-international.org/publications-and-standards/standards/ecma-262/.

Go FAIR (2016). FAIR Principles. https://www.go-fair.org/fair-principles/.

ISO (2025). The International Organization for Standardization. https://www.iso.org/.

Schema.org (2025). Schemas for structured data. https://schema.org/.

Wikipedia (2025). Open science. https://en.wikipedia.org/wiki/Open_science.

Central Limit Theorem for a Set-valued Martingale

Andrej Srakar[1,⋆]

1. Institute for Economic Research Ljubljana
⋆Contact author: andrej.srakar@ier.si

Keywords: Central limit theorem, Set-valued process, **F**-martingale

In classical data analysis, objects and patterns are usually described by a vector of qualitative or quantitative measurements, where each column represents a single variable. This is too restrictive to represent more complex data. To take into account the uncertainty and/or variability to the data, variables can assume sets of categories or intervals even with frequencies or weights. Such kind of data have been mainly studied in Symbolic Data Analysis (SDA) (see, for example, Bock and Diday, 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture, 2008; Noirhomme-Fraiture and Brito, 2011). While now with several developed approaches and work on foundations in terms of asymptotics, inference and frequentist/Bayesian alternatives (see e.g. Zhang et al., 2020), the field is still in its infancy with many standard methodological approaches lacking proper development.

In our article, we address the necessary tools for asymptotic analysis of symbolic data. It can be studied in the context of set-valued stochastic processes (see, e.g., Schmelzer, 2013). A random set is a random variable whose values are subsets of some set E. In the following, let E be a topological space and let $G(E)$ and $F(E)$ denote, respectively, the family of open and closed subsets of E. The Borel σ-algebra on E is denoted by $B(E)$ and $P(E)$ is the power set of E. Let (Ω, Σ, μ) be a σ-finite measure space. Let $A : \Omega \to P(E)$ be a set-valued map / multifunction in a measurable space (Ω, Σ). The map A is called weakly (or Offros) measurable if its upper inverse $A^-(G) \in \Sigma$ for all $G \in G(E)$, and it is called strongly measurable if $A^-(B) \in \Sigma$ for all $B \in B(E)$. For a weakly measurable, almost surely non-empty multifunction $A : \Omega \to F(E)$ we define $S^p(A) = \{\alpha \in L^p(\Omega; \Sigma) : \alpha(\omega) \in A(\omega)$ for almost all $\omega \in \Omega\}$ which is a closed subset of $L^p(\Omega; \Sigma)$. If $S^p(A) \neq 0$ there exists a Castaing representation consisting of integrable selections, that is a sequence $\{\alpha_n\}_{(n \in N)} \subseteq S^p(A)$ such that $A(\omega) = cl(\{\alpha_n(\omega)\}_{(n \in \mathbf{N})}$for all $\omega \in D(A))$.

We let X be a separable Banach space with norm $|\cdot|$. We shall denote $P(X)$ to be the set of all nonempty subsets of X, $C(X)$ to be the set of all closed sets of $P(X)$, and $K(X)$ to be the set of all compact convex sets in $P(X)$ with respect to the norm topology on X. Next, we let $L^0(E, C(X)) = L^0_\varepsilon(E, C(X))$ be the set of all measurable set-valued mappings $F : E \to C(X)$ distinguished up to μ-a.e. equality. A set-valued stochastic process $\Phi = \{\phi_t\}_{(t \in [0,T])}$ is a family of set-valued random variables taking values in $C(R^d)$. We call Φ measurable if it is $B([0, T]) \bigotimes F$-measurable as a single set-valued function on $[0, T] \times \Omega$. We denote $L^0_F([0, T] \times \Omega, C(R^d))$ to be the space of all set-valued, **F**-progressively measurable processes taking values in $C(R^d)$. We say that a set-valued process $M = \{M_t\}_{(t \in [0,T])}$ is a set-valued **F**-martingale if $M \in L^0_F([0, T] \times \Omega, C(R^d)), M_t \in A^1_{F_t}(\Omega, C(R^d)), 0 \leq t \leq T$.

In our article we use Stein approaches and Fourier analysis to prove the following central limit theorem for the set-valued **F**-martingale:

Theorem: Let $M = \{M_t\}_{t \in [0,T]}$ be a set-valued uniformly square-integrably bounded **F**-martingale. Let $H = \{H_t\}_{t \in [0,T]}$ be a set of bounded and predictable processes. Take a right-continuous filtra-

tion F_t. Let $U_i(t) = \int_0^t H_i(u)dM_i(u), i = 1, 2, ...$ We write $U(\cdot) = \int_0^\cdot f(s)dW(s)$, where $W(\cdot)$ is a Wiener process, and where $f(\cdot)$ is such that $\int_0^t f^2(s)ds = \alpha(t)$. Then $\Sigma U_n(\cdot) \longrightarrow^w U(\cdot)$, where $U(\cdot)$ is a zero-mean Gaussian process with independent increments and variance function $\alpha(\cdot)$.

The above central limit theorem can be proven also for more general set-valued quantities and processes and we discuss its extensions.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.

Bock, H.H. and Diday, E. (2000). *Analysis of Symbolic Data, Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin-Heidelberg.

Diday. E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, Chichester.

Noirhomme-Fraiture, M. and Brito, P (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining 4(2)*, 157–170.

Schmelzer, B. (2013). Set-valued assessments of solutions to stochastic differential equations with random set parameters. *J. Math. Anal. Appl. 400(2)*, 425–438.

Zhang, X. and Beranger, B. and Sisson, S.A. (2020). Constructing likelihood functions for interval-valued random variables. *Scandinavian Journal of Statistics 47(1)*, 1–35.

Multiblock Analysis of Distributional Data

Paula Brito[1,⋆], **Ndeye Niang**[2], **A. Pedro Duarte Silva**[3], **Stephanie Bougeard**[4]

1. Faculty of Economics, University of Porto & LIAAD INESC TEC, Portugal
2. CEDRIC, Conservatoire National des Arts et Métiers, Paris, France
3. Universidade Católica Portuguesa, Católica Porto Business School and CEGE, Portugal
4. Department of Epidemiology, ANSES, France

⋆Contact author: mpbrito@fep.up.pt

Keywords: Histogram data, Multiblock regression, PLS regression.

We are interested in distributional numerical data, where, for each variable, the units are described by empirical distributions. In our model, each distribution is represented by a location measure and interquantile ranges, for a chosen set of quantiles (Brito and Duarte Silva, 2025). This leads naturally to blocks of indicators associated with each of the descriptive variables. We propose to take into account this structure into homogeneous blocks using multiblock methods. Within this framework, both supervised as well non-supervised methods may be addressed.

In this work, we address regression between distribution-valued variables within the multiblock framework. As the indicators in each block may, by nature, be strongly related, we use PLS multiblock regression to manage multicollinearity (Bougeard and Dray, 2018). In our model, there is one response block corresponding to the distributional-valued target variable Y, with the respective indicators, and several similar blocks corresponding each to a descriptive distributional-valued variable X_j, all blocks having the same size. The method allows understanding the importance in the model of each distributional variable as a whole (interpretation at the block-level), as well as the relevance of each individual indicator of each descriptive variable (interpretation at the indicator-level).

Applications to real data put in evidence the interest of the proposed approach.

Acknowledgments
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 https://doi.org/10.54499/LA/P/0063/2020, and UID/GES/00731/2019.

References

Bougeard, S., & Dray, S. (2018). Supervised Multiblock Analysis in R with the ade4 Package. *Journal of Statistical Software 86*, 1–17.

Brito, P., & Duarte Silva, A. P. (2025). Parametric Models for Distributional Data. *Advances in Data Analysis and Classification*, in press.

Geometric Goodness-of-fit Measure for Interval-valued Data

Dylan Benavides[1,2,⋆], **Oldemar Rodríguez**[2], **Maikol Solís**[2]

1. National Center for High Technology
2. CIMPA, School of Mathematics, University of Costa Rica
⋆Contact author: dbenavides@cenat.ac.cr

Keywords: Goodness-of-fit, Interval-type variables, Rectangular clusters, Alpha shapes, Empty space loss function.

Statistical and data analysis methods have been developed mainly in cases where variables take a single value, however, there are several situations in which the use of this type of variables can cause the loss of relevant information. In the case of quantitative variables, more complete information can be achieved by describing a set of statistical units as interval-type symbolic data, that is, when the value taken by a variable is an interval of the form $[a, b]$, with $a, b \in \mathbb{R}$, $a \leq b$.

Using this type of data offers computational advantages, since it is possible to summarize large data sets in a more manageable size, keeping as much information as possible from the original database. Different methods have been developed for the analysis of interval-type symbolic data, such as regression models, principal component analysis, correspondence analysis, among others.

In symbolic linear regression models with the method of centers and ranges, one of the most commonly used measures of goodness of fit is the symbolic coefficient of determination, which is an extension of the coefficient of determination in classical data, because a classical regression model is applied to the centers matrix and another regression model to the ranges matrix Lima Neto, E. and De Carvalho, F. (2010).

In the article called *Geometric goodness of fit measure to detect patterns in data point clouds*, Hernández, A. and Solís, B. (2023) present the construction of a geometric goodness of fit index for classical data, similar to the coefficient of determination R^2.

The use of alpha shapes, which represent a continuous extension of the data-point set, is applied to point clouds that are obtained from the results of a regression model, and it is established that the index measures the difference in area between the alpha shape and the smallest rectangular window that contains the cloud of data points.

This index has proven to be very useful for the study of regression models, since it allows the identification of geometric patterns that relate the predictor variables to the response variable, likewise, in the article *UMAP projections and the survival of empty space: A geometric approach to high-dimensional data* Solís, M. and Hernández, A. (2024), this index was applied to a dimensionality reduction problem.

As in classical data, in symbolic data the coefficient of determination has some shortcomings; R^2 increases if new variables are added to the model, in small samples it tends to be larger and ignores the geometric arrangement of the data Cramer, J. (1987), so now the interest is to be able to establish a geometric goodness of fit index for symbolic data of the interval-type that solves these problems.

This work extends the concept of geometric goodness-of-fit index to interval-type symbolic data, focusing exclusively on the geometric characteristics of point clouds associated with rectangular

clusters. The proposed index: $R^2_{Int,\alpha}$ is based on the analysis of the empty space loss function using alpha shapes, allowing for the identification of geometric structures in these clusters.

The study examines how the index evolves with respect to the parameter α, demonstrating its capacity to detect relevant geometric features in symbolic regression models. Two algorithms are proposed to compute the index: *Method of centers* and *Method of centers and extremes* and their effectiveness is assessed through experiments using symbolic linear regression models applied to interval-type data.

Results indicate that the proposed index captures global geometric characteristics and reveals internal and external structures in the data. Furthermore, the method of centers and extremes offers an improvement over the method of centers by considering the variability of interval-type variables.

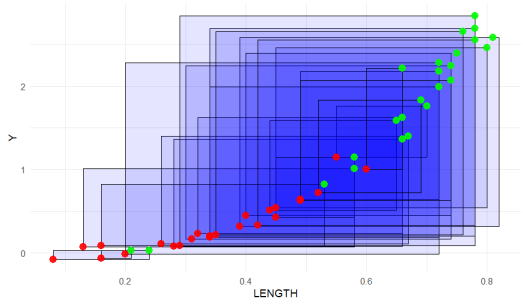


Figure 1: Left and Right extremes of a rectangular clustering

Figure 1 shows one of the stages of the developed algorithm, in which two point clouds are extracted from a rectangular cluster associated with a predictor variable of a regression model.

References

Cramer, J. (1987). Mean and variance of R^2 in small and moderate samples. *Journal of Econometrics 35*, 253–266.

Lima Neto, E., De Carvalho, F. (2010). Constrained linear regression models for symbolic Interval-valued variables. *Computational Statistics and Data Analysis 54*, 333–347.

Edelsbrunner, H. (2014). *A Short Course in Computational Geometry and Topology*. Springer

Lima Neto, E., De Carvalho, F. (2017). Nonlinear regression applied to Interval-valued data. *Pattern Analysis and Applications 20*, 809–824.

Arce, J., Rodríguez, O. (2019a). Optimized dimensionality reduction methods for Interval-valued variables and their application to facial recognition. *Entropy 21*, 1016.

Baddeley, A., Rubak, E., Turner, R. (2019b). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, New York

Chacón, J., Rodríguez, O. (2021). Regression models for symbolic Interval-valued variables. *Entropy 23*, 429.

Hernández, A., Solís, B. (2023). Geometric goodness of fit measure to detect patterns in data point clouds. *Computational Statistics 38*, 1231–1253.

Solís, M., Hernández, A. (2024). Umap projections and the survival of empty space: A geometric approach to high-dimensional data. In *Proceedings of the 18th Conference of the International Federation of Classification Societies Conference*, pp. 1–9.

Dimensionality Reduction

Riemannian Principal Component Analysis for Interval-valued Data

Oldemar Rodríguez[1,⋆]

1. CIMPA, School of Mathematics, University of Costa Rica
⋆Contact author: oldemar.rodriguez@ucr.ac.cr

Keywords: Riemannian manifold, Riemannian principal component analysis (R-PCA), Riemannian statistics, Local distance notion, Dimension reduction geometric structures, Riemannian statistics

This paper proposes an innovative extension of Principal Component Analysis (PCA) that transcends the traditional assumption of data lying in Euclidean space, enabling its application to data on Riemannian manifolds. The primary challenge addressed is the lack of vector space operations on such manifolds. Fletcher et al., in their work *Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape*, see Fletcher et al. (2004), proposed Principal Geodesic Analysis (PGA) as a geometric approach to analyze data on Riemannian manifolds, particularly effective for structured datasets like medical images, where the manifold's intrinsic structure is apparent. However, PGA's applicability is limited when dealing with general datasets that lack an implicit local distance notion. In this paper, we introduce a generalized framework, termed *Riemannian Principal Component Analysis (R-PCA)*, to extend PGA for any data endowed with a local distance structure. Specifically, we adapt the PCA methodology to Riemannian manifolds by equipping data tables with local metrics, enabling the incorporation of manifold geometry. This framework provides a unified approach for dimensionality reduction and statistical analysis directly on manifolds, opening new possibilities for datasets with region-specific or part-specific distance notions, ensuring respect for their intrinsic geometric properties. In the paper Rodríguez (2025), you can see the detail. We, also, examine how R-PCA significantly improves upon the Center Method in Principal Component Analysis for Interval-valued data.

References

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine 29*(6), 141–142.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging 23*(8), 995–1005. https://doi.org/10.1109/TMI.2004.831793.

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., and Sanes, J. R. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell 166*(5), 1308–1323.e30. https://doi.org/10.1016/j.cell.2016.07.054.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (Nov), 2579–2605.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* arXiv:1802.03426. http://arxiv.org/abs/1802.03426.

Oudot, S. Y. (2016). *Persistence Theory: From Quiver Representations to Data Analysis*. Mathematical Surveys and Monographs, Vol. 209. American Mathematical Society.

Pennec, X., Sommer, S., and Fletcher, T. (Eds.). (2020). *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press, Elsevier.

Rabadán, R., and Blumberg, A. J. (2020). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press.

Rodríguez, O. (2025). Riemannian Statistics for Any Type of Data. *In: Trejos, J., Chadjipadelis, T., Grané, A., Villalobos, M. (eds) Data Science, Classification, and Artificial Intelligence for Modeling Decision Making. IFCS 2024. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Cham. https://doi.org/10.1007/978-3-031-85870-3_17

Principal Component Analysis of Distributional Data

Sónia Dias[1,2,⋆], **Paula Brito**[1,3]

1. LIAAD - INESC TEC
2. Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Viana do Castelo, Portugal
3. Faculty of Economics, University of Porto, Portugal
⋆Contact author: sdias@estg.ipvc.pt

Keywords: Histogram-valued variables, Principal component analysis, DSD regression model.

Currently, the development of models and methods for the representation, analysis, interpretation and organization of distributional data is growing (Brito, P., Dias, S., 2022). Linear models are the basis of several statistical methods, such as linear regression, linear discriminant analysis and principal component analysis. The Distribution and Symmetric Distribution (DSD) linear regression model proposed in Dias S. and Brito P. (2015) allows predicting the distribution of the target variable from other histogram-valued variables, and is obtained optimizing a criterion based on the Mallows distance between the observed and the predicted distributions.

In this work a Principal Component Analysis that uses the definition of linear combination considered in the DSD Model is proposed. Each principal component is obtained by a linear combination of the p original correlated histogram-valued variables as follows:

$$\Psi_\epsilon(t) = \sum_{j=1}^{p} a_j \Psi_{X_j}(t) - b_j \Psi_{X_j}(1-t) \quad \text{with} \quad a_j, b_j \geq 0$$

where $\Psi_{X_j}(t)$ and $\Psi_{X_j}(1-t)$ represent, for each individual, the quantile function of the histogram X_j and the quantile function of the respective symmetric histogram, respectively.

In this work, we consider the definition of covariance between histogram-valued variables as proposed by Irpino A. and Verde R. (2015), and which is based on the Mallows distance.

For the first principal component, the vector of the non-negative parameters $\gamma = [a_1\, b_1\, \ldots\, a_p\, b_p]$ is estimated maximizing the variance of the first principal component, that in this case is a quantile function, $\Psi_{\epsilon_1}(t)$. The definition of variance (Irpino A. and Verde R. , 2015) is the follows:

$$\begin{aligned} var(\Psi_{\epsilon_1}(t)) &= \frac{1}{n}\sum_{i=1}^{n} D_M^2\left(\Psi_{\epsilon_1(i)}(t), \overline{\Psi_{\epsilon_1}}(t)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{m} p_\ell \left[\left(c_{\epsilon_1(i)_\ell} - \overline{c}_{\epsilon_1\ell}\right)^2 + \frac{1}{3}\left(r_{\epsilon_1(i)_\ell} - \overline{r}_{\epsilon_1\ell}\right)^2\right] \end{aligned}$$

where $\overline{\Psi_{\epsilon_1}}(t)$ is the barycenter of $\Psi_{\epsilon_1(i)}(t)$ and $\overline{c}_{\epsilon_1\ell}$, $\overline{r}_{\epsilon_1\ell}$ are the mean of the centers and of the half-ranges of the sub-intervals, respectively.

Similarly to the classical statistics but considering the definitions presented above, the parameters for the first principal component are obtained.

$$Maximize \quad var(\Psi_{\epsilon_1}(t))$$

$$subject\,to$$

$$\gamma\gamma^T = 1$$

$$a_j, b_j \geq 0$$

Maximisation of the variance of the first principal component is obtained by solving this quadratic optimization problem.

The proposed approach for determining the first principal component may be particularized to Interval-valued variables, which constitute a special case of histogram-valued variables.

Examples illustrate the behaviour of the method in different situations.

References

Brito, P., Dias, S. (2022) *Analysis of Distributional Data*. CRC Press, Taylor & Francis Group.

Dias S. and Brito P. (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining 8(2)*, 75–113.

Irpino A. and Verde R. (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification 9(2)*, 143–175.

Foundations of SDA and Statistical Modelling II

An Extension of Entropy for Interval-valued Data

José Andrés Piedra-Molina[1,⋆], **Oldemar Rodríguez**[1,⋆]

1. CIMPA, School of Mathematics, University of Costa Rica
⋆Contact author: jose.piedramolina@ucr.ac.cr, oldemar.rodriguez@ucr.ac.cr

Keywords: Symbolic data analysis, Interval-valued data, Entropy, Combinatorial optimization.

In symbolic data analysis the objects of study are second-order units, that is data resulting either from aggregation or observation, where there is inherent uncertainty regarding the precise measurements of each first-order unit. We propose the adaptations of entropy in information theory–δ-information content and δ-entropy–to quantify uncertainty in the case of interval–valued data. We define the δ-information content of a non-degenerate interval L as $\eta_\delta(L) = -\log \frac{\delta}{m(L)}$ and propose the δ-entropy of an Interval-valued random variable Z with probability distribution $p(Z_i) = p_i$ as its expected δ-information content, i.e. $H_\delta(Z) = E(\eta_\delta(Z)) = -\sum_i p_i \log \frac{\delta}{m(Z_i)}$. We show that δ-entropy enjoys several analogous properties to regular entropy and that it can measure two important layers of uncertainty in Interval-valued data: the distribution of symbolic data itself, and the distribution of first-order units within their respective symbolic representations. We conclude with an application to select the concepts of minimum δ-entropy with which to represent real-valued data as Interval-valued data.

References

Hu, C., Hu, Z. (2020). On statistics, probability, and entropy of Interval-valued datasets. *Information Processing and Management of Uncertainty in Knowledge-Based Systems 7*, 407–421.

Advancements in the Best Point Method: New Optimization Criteria

Mario J. Gómez[1], **Jorge Arce**[2], **Oldemar Rodríguez**[3,⋆]

1. Graduate Program in Mathematical Methods and Applications, University of Costa Rica
2. School of Mathematics, National University
3. CIMPA, School of Mathematics, University of Costa Rica

⋆Contact author: oldemar.rodriguez@ucr.ac.cr

Keywords: Principal components, Interval data, Best point method.

Since the first exploration of PCA for interval data with the centers method (Cazes et al., 1997), where the Interval-valued matrix is reduced to midpoints for traditional PCA, several alternative approaches have emerged. For example, the vertices method (Cazes et al., 1997) transforms intervals into matrices of vertices treated as independent observations, although this approach was later criticized by Chouakria et al. (2011) for inadequately capturing internal variability and for the inherent dependence among vertices. In response to these issues, Lauro and Palumbo (2000) advanced the field with methods such as symbolic-object PCA, range transformation PCA, and a mixed strategy, and further refined the approach with the midpoint-radii method (Palumbo and Lauro, 2003) by separating centers and ranges in the analysis.

Building on these developments, more sophisticated methods have been proposed. Gioia and Lauro (2006) introduced an interval algebra-based PCA, while other approaches like the Complete-Information-based PCA (CIPCA) (Wang et al., 2012) and the Symbolic Covariance PCA (SC-PCA) (Le-Rademacher and Billard, 2012) have been developed, offering alternative mathematical formulations to handle interval data. As one of the latest advancements in this area, the Best Point (BP) method (Arce and Rodríguez, 2019) challenges the assumption that the center is the optimal representative for PCA analysis by instead strategically selecting an optimal point within each interval based on specific optimization criteria.

Although the original BP method yielded valuable insights, it was limited in its optimization scope. Building on its inherent flexibility, we propose an extended analytical framework that incorporates three additional optimization criteria: individual representation, variable representation, and angles between variables.

To address the challenge of balancing these potentially competing objectives, we adopt the multi-objective optimization framework developed by Monteil et al. (2020). This approach integrates genetic algorithms (Eiben and Smith, 2015) with an enhanced version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Liu and Nocedal, 1989) to simultaneously optimize across all five criteria.

The practical benefits of this enhanced BP approach are illustrated through a case study using the face recognition dataset by Leroy et al. (1996), which comprises Interval-valued features derived from images, capturing variability across multiple instances. Comparing the performance of the traditional center method, vertex method, and the five BP variants, the results show substantial improvements in the new optimization criteria. Notably, the multi-objective approach achieves a balanced compromise among competing goals, although some trade-offs remain.

Empirical findings highlight the method's capacity to preserve the primary data structure while of-

fering deeper insight into the internal variability of symbolic observations. Visualizations demonstrate that while the overall configuration of principal components remains consistent across methods, the dispersion and coverage of the intervals vary significantly. This indicates that the choice of optimization criteria can profoundly influence the interpretability and utility of PCA outcomes in symbolic contexts.

The enhancements to the BP method provide researchers with a more flexible and powerful tool for analyzing complex Interval-valued data. This work also opens new directions for future research, particularly in refining optimization techniques and exploring broader applications across domains.

References

Arce, J., Rodríguez, O. (2019). Optimized Dimensionality Reduction Methods for Interval-valued Variables and Their Application to Facial Recognition. *Entropy 21 (10)*, 1016.

Cazes, P., Chouakria, A., Diday, E., Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique appliquée 45 (3)*, 5–24.

Chouakria, A., Billard, L., Diday, E. (2011). Principal component analysis for Interval-valued observations. *Statistical Analysis and Data Mining: The ASA Data Science. Journal 4 (2)*, 229–246.

Eiben, A. E., Smith, J. E. (2015). *Introduction to evolutionary computing.* Springer Berlin, Heidelberg.

Gioia, F., Lauro, C. (2006). Principal component analysis on interval data. *Computational statistics 21 (1)*, 343–363.

Lauro, C., Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational statistics 15 (1)*, 73–87.

Le-Rademacher, J., Billard L. (2012). Symbolic Covariance Principal Component Analysis and Visualization for Interval-valued Data. *Journal of Computational and Graphical Statistics 21 (2)*, 413–432.

Leroy, B., Chouakria, A., Herlin, I., Diday, E. (1996). Approche géométrique et classification pour la reconnaissance de visage. In *Proceedings of the Congrés de Reconnaissance des Formes et Intelligence Artificielle (Rennes, France).* pp. 548–557.

Liu, D. C., Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming 45 (1)*, 503–528.

Monteil, C., Zaoui, F., Le Moine, N., Hendrickx, F. (2020). Multi-objective calibration by combination of stochastic and gradient-like parameter generation rules – the caRamel algorithm. *Hydrology and Earth System Sciences 24 (6)*, 3189–3209.

Palumbo, F., Lauro, C. (2003). A PCA for Interval-valued data based on midpoints and radii. In *Proceedings of the International Meeting of the Psychometric Society IMPS2001 (Osaka, Japan).* pp. 641–648.

Wang, H., Guan R., Wu J. (2012). CIPCA: Complete-Information-based Principal Component Analysis for Interval-valued data. *Neurocomputing 86 (1)*, 158–169.

Symbolic Data as Matrix-valued Data Robust Estimation and Explainable Outlier Detection

Marcus Mayrhofer[1⋆], **Paula Brito**[2⋆], **A. Pedro Duarte Silva**[3], **Peter Filzmoser**[1]

1. Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria
2. Faculty of Economics, University of Porto & LIAAD INESC TEC, Portugal
3. Universidade Católica Portuguesa, Católica Porto Business School and CEGE, Portugal
⋆Contact author: mpbrito@fep.up.pt

Keywords: Histogram data, Matrix-valued data, Explainable AI, Shapley values.

We consider numerical distributional data, where units are described by histogram or Interval-valued variables $Y_j, j = 1, \ldots, p$. In our model, each distribution is represented by a central statistic C, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles $\phi_1, \ldots, \phi_q$, denoted $R^*_h, h = 1, \ldots, m$, where $m = q - 1$ is the number of considered intervals. Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles (Brito and Duarte Silva, 2025); Interval-valued data are represented by midpoints and log-ranges ($m = 1$) (Brito and Duarte Silva, 2012).

Multivariate Normal distributions are then assumed for the whole set of indicators. Furthermore, we consider alternative structures of the variance-covariance matrix. In the most general formulation we allow for non-zero correlations among all central statistics and log-ranges; for distributional variables there are however other cases of interest, whether the variables, the central statistics and the different log-ranges, are or are not correlated between or among themselves, leading to five different configurations in the distributional data case, and four configurations for Interval-valued data.

In this work we consider these data as matrix-valued data, represented as a tensor of dimension $n \times p \times q$, following (Mayrhofer et al, 2025), $X \sim ME(M, \Sigma_{var}, \Sigma_{ind}, g)$, where

- Σ_{var} is $p \times p$ and gathers variances and covariances between the variables Y_j
- Σ_{ind} is $q \times q$ and gathers variances and covariances between the considered indicators $C, R^*_1, \ldots, R^*_m$
- $g(z) = \dfrac{\exp(-z/2)}{2\pi^{pq/2}}$

In this model, the global covariance matrix Σ is written as $\Sigma = \Sigma_{ind} \otimes \Sigma_{var}$. This implies that we assume that covariances between the different indicators are constant across variables, thereby obtaining a more parsimonious model and reducing the number of parameters to be estimated. The different covariance configurations correspond to setting Σ_{ind} and/or Σ_{var} as block-diagonal matrices.

The Matrix Minimum Covariance Determinant (MMCD) method (Mayrhofer et al, 2025) accounts for the matrix-variate data structure and robustly estimates the mean matrix, as well as the row-wise Σ_{var} and column-wise Σ_{ind} covariance matrices.

Robust Mahalanobis distances based on MMCD estimators then allow for outlier detection. Using the concept of Shapley values for outlier explanation in this the matrix-variate setting, enables the

decomposition of the squared Mahalanobis distances into contributions of the variables, indicators, and individual cells of the matrix-valued observations.

Applications to real data put in evidence the interest of the proposed approach.

Acknowledgments
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 https://doi.org/10.54499/LA/P/0063/2020, and UID/GES/00731/2019.

References

Mayrhofer, M., Radojičić, U., Filzmoser, P. (2025). Robust covariance estimation and explainable outlier detection for matrix-valued data. *Technometrics*, in press.

Brito, P., Duarte Silva, A. P. (2025). Parametric models for distributional data. *Advances in Data Analysis and Classification*, in press.

Brito, P., Duarte Silva, A. P. (2012). Modelling interval data with Normal and Skew-normal distributions. *Journal of Applied Statistics, 39(1)*, 3-20.

Explainable Outlier Detection in Interval-valued Data Using a Robust Covariance Estimator

Catarina P. Loureiro[1,2,*], **M. Rosário Oliveira**[1,2], **Paula Brito**[3,4], **Lina Oliveira**[2,5]

1. CEMAT
2. Dep. Mathematics, Instituto Superior Técnico, Lisbon, Portugal
3. LIAAD-INESC TEC
4. Fac. Economia, Universidade do Porto, Porto, Portugal
5. CAMGSD

*Contact author: catarinapadrela@tecnico.ulisboa.pt

Keywords: Robust statistics, Minimum covariance determinant estimator, Mallows' distance, Outlier detection, Shapley value

Symbolic Data Analysis focuses on modelling complex data structures, preserving the data's underlying variability while also reducing the dataset size. In this work, we concentrate on one of the most commonly used symbolic data types, Interval-valued data. These data structures can be useful when dealing with Big Data, but not without presenting challenges that call for novel strategies.

One of these challenges is estimating the location and scale of Interval-valued random vectors. This can be done using the barycentre approach based on the Mallows' distance (Irpino and Verde, 2015; Oliveira et al., 2024). Nevertheless, the presence of anomalous data points in real-life datasets raises an additional issue, since it can strongly influence these (classical) estimates, leading to the necessity of robust methods. To address this, we extend the Minimum Covariance Determinant (MCD) estimator (Rousseeuw and van Driessen, 1999) to Interval-valued data in order to obtain robust estimators of location and scale. The first step is to define the optimization problem and proving the concavity of the objective function (Boyd and Vandenberghe, 2004). Under these conditions, the MCD estimator for Interval-valued data can be derived applying the Majorization-Minorization algorithm (Lange, 2016) with Taylor's expansion.

As a product of the MCD algorithm, we obtain a robust distance that can be used in detecting outlier observations. As in the conventional case, outlier detection can be accomplished by assigning suitable cut-off values or even by exploiting the farness concept (Raymaekers et al., 2022). Additionally, this robust distance can be decomposed into each variable's outlyingness contribution, using the Shapley value, a game theory concept that has become prominent in Explainable AI. This contributes to the interpretation of Interval-valued multivariate outliers, following the idea of the Mahalanobis distance decomposition in conventional statistics (Mayrhofer and Filzmoser, 2023).

In the interest of evaluating the performance of the Interval-valued MCD estimator and outlier detection method, a simulation study is conducted. We compare the proposed robust estimator with the classical estimators across several contamination levels, using synthetically generated symbolic datasets. The results show that, for all considered levels of contamination, the Interval-valued MCD estimator consistently outperforms its classical counterpart in estimating the symbolic covariance matrix. As for the outlier detection method, it achieves high accuracy, specially when paired with the farness concept.

Finally, we apply our outlier detection method to a real-life dataset, where the Shapley values prove to be particularly helpful in interpreting the identified outlier observations. Our findings

show that the proposed methods are powerful tools for robust symbolic data analysis in real-world applications.

Acknowledgments
We thank FCT - Fundação para a Ciência e Tecnologia for the grant UI/BD/153720/2021, and the projects UIDB/04621/2020, UIDB/04459/2020, UIDP/04459/2020, and LA/P/0063/2020.

References

Irpino, A., Verde, R. (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification 9(2)*, 143–175.

Oliveira, M.R., Pinheiro, D., Oliveira, L. (2024). Location and association measures for interval data based on Mallows' distance. https://arxiv.org/abs/2407.05105.

Rousseeuw, P.J., van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics 41(3)*, 212–223.

Boyd, S., Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Lange, K. (2016). *MM Optimization Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia.

Raymaekers, J., Rousseeuw, P.J., Hubert, M. (2022). Class Maps for Visualizing Classification Results. *Technometrics 64(2)*, 151–165.

Mayrhofer, M., Filzmoser, P. (2023). Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*, 2452–3062.

Clustering

Clustering Density-valued Data

Rui Nunes[1,2,⋆], **Paula Brito**[1,3], **Sónia Dias**[1,4]

1. LIAAD - INESC TEC
2. Faculdade de Ciências - Universidade do Porto
3. Faculdade de Economia - Universidade do Porto
4. Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Viana do Castelo

⋆Contact author: up201400313@up.pt / rui.miguel.nunes@inesctec.pt

Keywords: Clustering, Symbolic data, Similarities, Density-valued data, KDE.

In this study, we are concerned with multivariate numerical distributional data. Recent studies have focused on histogram-valued variables. Each unit i of a histogram-valued variable Y can be represented by the classical representation of a histogram $H_{Y(i)}$ or its corresponding quantile function $\Psi_{Y(i)}(t)$ with $t \in [0,1]$ under some distributional assumption.

We focus on density-valued variables, instead of histogram-valued variables, where each unit i is represented by its (estimated) density or the corresponding quantile function. One can estimate the density-valued variables with a non-parametric kernel density estimator (KDE). Considering that each unit i corresponds to N observations at microdata level, we can express KDE as

$\hat{f}(x) = \frac{1}{Nh}\sum_{k=1}^{N} K\left(\frac{x-X_k}{h}\right)$, where h is the bandwidth and $K(u)$ is a kernel function; the Gaussian kernel is often used and can be expressed by $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$.

Clustering aims at identifying groups of similar units within a dataset. This study focuses on hierarchical clustering, particularly agglomerative approaches. Dissimilarity measures are pivotal for density-based clustering. Table 1 lists the dissimilarity measures used in our study.

Table 1: Dissimilarity measures between density functions.

Bhattacharyya	$D_B(f_1(x), f_2(x)) = -ln\left(\int_{\mathcal{X}} \sqrt{f_1(x)f_2(x)}dx\right)$
Hellinger	$D_H(f_1(x), f_2(x)) = \sqrt{1 - \int_{\mathcal{X}} \sqrt{f_1(x)f_2(x)}dx}$
Mallows	$D_M(\Psi_1(t), \Psi_2(t)) = \sqrt{\int_0^1 (\Psi_1(t) - \Psi_2(t))^2 dt}$
Total Variation	$D_{TV}(f_1(x), f_2(x)) = \frac{1}{2}\int_{\mathcal{X}} \lvert f_1(x) - f_2(x)\rvert dx$
Jeffreys divergence	$D_J(f_1(x), f_2(x)) = \int_{\mathcal{X}} (f_1(x) - f_2(x)) ln\big(f_1(x)/f_2(x)\big)dx$
L^2	$D_{L^2}(f_1(x), f_2(x)) = \sqrt{\int_{\mathcal{X}} (f_1(x) - f_2(x))^2 dx}$

In Table 1, $f_1(x), f_2(x)$ are density functions where we assume the same domain $\mathcal{X}$ and $\Psi_1(t), \Psi_2(t)$ are the quantile functions of density $f_1(x)$ and $f_2(x)$ respectively, with $t \in [0,1]$. In hierarchical clustering methods, the choice of linkage criteria plays a crucial role in defining the clustering structure. For this study, we have considered the Single, Complete, and Unweighted average (UPGMA) Linkage methods described in Table 2.

Table 2: Linkage methods.

Single linkage	$D^C(A,B) = \min_{a\in A, b\in B} D^U(a,b)$
Complete linkage	$D^C(A,B) = \max_{a\in A, b\in B} D^U(a,b)$
Unweighted average linkage	$D^C(A,B) = \frac{1}{\lvert A\rvert.\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} D^U(a,b)$

In Table 2 A and B represent two clusters, and $D^U(a,b) = \sqrt{\sum_{j=1}^{p} D(a_j,b_j)^2}$ denotes the distance between units, where each unit is described by p variables, and distance D is defined in Table 1.

The method was applied to a dataset of 31 European countries' GDP between 1995 and 2022 and includes the following GDP components "Private Consumption, "Public Consumption", "Gross Capital Formation", "Export Goods" and "Import Goods". To decide on the number of clusters, we use the well-known Silhouette coefficient described by Peter J. Rousseeuw (1987). Usually, a higher value of the silhouette coefficient means a better separation. In addition to the value itself, one can visualize the silhouette plot and validate its structure. The clustering allowed us to put in evidence groups of countries with similar distributions of the variables considered. In Figure 1

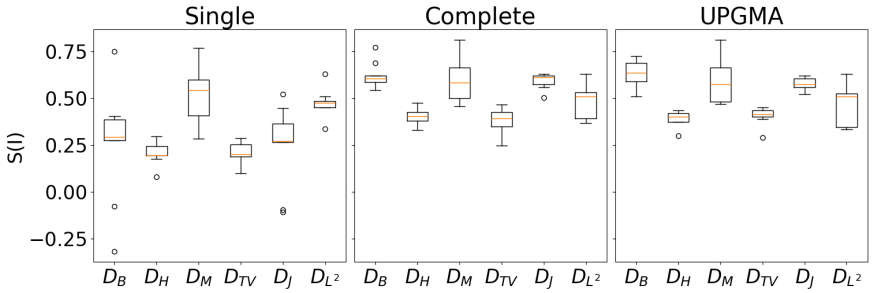


Figure 1: Silhouette coefficient when k varies between 2 and 10

we can observe the variability of the silhouette coefficient across different numbers of clusters ($k = 2, \cdots, 10$) for different linkage methods and dissimilarity measures. One can argue that the Complete and UPGMA linkages provide the best silhouette coefficient on average, in particular for the *Jeffreys divergence*, *Bhattacharyya* and *Mallows* dissimilarity measures, and that Single linkage has the worst performance. A simulation study was performed for these dissimilarity measures to further investigate their behaviour in this context.

Acknowledgements
This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020, DOI 10.54499/LA/P/0063/2020.

References

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: the basic choices. In *Proceedings of IFCS (1987), Classification and Related Methods of Data Analysis (Aachen, Germany)*, pp. 673–684, North Holland.

Dias, S., Brito, P. (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal 8(2)*, 75–113.

Rousseeuw, P. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53-65.

Clustering Intervals Using Principal Components

Jiankun Zhu[1], **Lynne Billard**[1⋆]

1. Lilly and University of Georgia
⋆Contact author: lynne@stat.uga.edu

Keywords: Interval data, Divisive clustering, Principal component analysis, Endpoints criteria.

Interval observations are examples of symbolic data. This article adapts the concepts of principal component analysis and clustering algorithms to develop a new methodology for interval data. Chavent's (1998, 2000) monothetic divisive clustering algorithm has been used extensively for clustering Interval-valued observations. To overcome some limitations of this algorithm, three new algorithms are proposed herein, one using the Chavent center based ordering idea but applied to principal components of each hypercube, one as a double ordering criteria using both interval endpoints, and a third as a mixed-strategy double algorithm that is based on the principal components criteria applied to both interval endpoints. Simulations show the proposed algorithms outperform previous methods; a real data set is analysed.

References

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters 19* , 989–996.

Chavent, M. (2000). Criterion-based divisive clustering for symbolic dat. . In: H.-H. Bock and E. Diday (Eds.), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* , Springer Berlin, pp. 299-–311.

Clustering Distributional Data Using Mahalanobis-like Distance on LDQ Functions

Gianmarco Borrata[1,2⋆], **Rosanna Verde**[1], **Antonio Balzanella**[1]

1. University of Campania *Luigi Vanvitelli*
2. University of Naples *Federico II*
⋆Contact author: gianmarco.borrata@unina.it

Keywords: Distributional-valued data, Logarithm derivative quantile functions, Clustering, Mahalanobis distance.

In today's big-data era, summarizing complex datasets with minimal information loss is crucial. Modern approaches often deal with numeric observations represented by probability or frequency distributions, capturing key features like mean, variability, and skewness of the underlying phenomenon. This gives rise to distribution-valued data (DD), where each datum is a univariate distribution. Within the Symbolic Data Analysis (SDA) framework Bock and Diday (2000), such variables are known as distributional variables, with histogram valued variables being a notable example. SDA extends classical methods (e.g., clustering, PCA, decision trees) to handle symbolic data such as intervals, categories, or distributions and is closely linked to multivariate analysis, and pattern recognition (Bock and Diday (2000)). Various clustering methods have been proposed for distribution-valued data, often grounded in Dynamic Clustering (DC) or k-means techniques. The dynamic clustering algorithm proposed by Diday and Simon (1976) is an iterative two-step method that alternates between forming clusters and identifying optimal prototypes (e.g., means, factorial axes, probability distributions) by locally minimizing an adequacy criterion.

Although several clustering methods for symbolic data have been proposed in the literature, few studies have explored the use of Mahalanobis distance for distributional variables. The extensions developed for interval data (Souza et al., (2004); De Carvalho and Lechevallier (2009)) do not generalize well to full distributions, mainly because density, quantile, and cumulative functions do not naturally lie in a Hilbert space, making the application of standard clustering techniques challenging.

To overcome this limitation, we propose a new dynamic clustering algorithm based on Mahalanobis distance for distributional data. The approach relies on two complementary components: the Log Derivative Quantile (LDQ) transformation (Petersen and Müller (2016)), which allows the shape and variability of distributions to be represented in a Hilbert space, and the minimum value of the quantile function, for recovering the location information lost during derivation. This combination enables consistent comparison of distributions in terms of shape, variability, and position, thereby enhancing clustering performance.

Let $\Omega = \{1, \ldots, n\}$ be a set of n objects described by p variables. Each object e_i $(i = 1, \ldots, n)$ is represented as a vector of LDQ functions $\mathbf{x}_i^l(t) = (x_i^{1l}(t), \ldots, x_i^{pl}(t))$ and a vector of minimum values $\mathbf{x}_i^m = (x_i^{1m}, \ldots, x_i^{pm})$. We assume that the prototype of a cluster C_k $(k = 1, \ldots, K)$ is also represented as a vector of average LDQ functions $\mathbf{g}_k^l(t)=(g_k^{1l}(t), \ldots, g_k^{pl}(t))$ and average scalar values $\bar{\mathbf{x}}_k^m=(\bar{x}_k^{1m}, \ldots, \bar{x}_k^{pm})$.

The dynamic clustering algorithm combines the Mahalanobis distance computed on the LDQ functions with the Mahalanobis distance computed on the minimum values. This allows the method to

account for both variability in the functional form and positional information:

$$J = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \left[d^2_{M_k^l}(\mathbf{x}_i^l(t), \mathbf{g}_k^l(t)) + d^2_{M_k^m}(\mathbf{x}_i^m, \bar{\mathbf{x}}_k^m) \right] \tag{1}$$

Dynamic clustering algorithms set an initial partition and alternate three steps until convergence, when the criterion J reaches a stationary value representing a local minimum.

We apply the algorithm to cluster Italian municipalities based on daily pollutant data (CO_2, PM10, PM2.5). Our objective is to identify clusters of municipalities that exhibit similar distributional pollutant profiles. The results highlight substantial differences in the distributions of pollutant levels in Italian municipalities. Specifically, spatial patterns reveal that municipalities in northern Italy exhibit systematically higher levels of emissions and particulate concentrations, while those in southern Italy show the lowest levels across all pollutants. Municipalities in the central regions display intermediate distribution profiles, reflecting a more mixed environmental context. The work opens up for further development and is a contribution in clustering and distributional data analysis techniques.

Acknowledgement
This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

Bock, H.-H., Diday, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer.

Diday, E. and Simon, J.C. (1976). *Digital pattern recognition.* Springer.

Souza, R.M.C.R., de Carvalho, F.A.T., Tenório, C.P., Lechevallier, Y. (2004). Dynamic cluster methods for interval data based on Mahalanobis distances. *Classification, Clustering, and Data Mining Applications*, 351–360.

de Carvalho, F.A.T., Lechevallier, Y. (2009). Dynamic clustering of Interval-valued data based on adaptive quadratic distances. *CIEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 1295–1306.

Petersen, A., Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics 44*, 183–218.

Regression

From Ordinary Least Squares to Robust Methods: Revisiting Regression for Interval-valued Variables

M. Rosário Oliveira[1,⋆], **Conceição Amado**[1]

1. CEMAT & Department of Mathematics, Instituto Superior Técnico, Lisbon, Portugal
⋆Contact author: rosario.oliveira@tecnico.ulisboa.pt

Keywords: Symbolic data analysis, Mallows' distance, Moore's algebraic structure, M-estimators

The Symbolic Data Analysis (SDA) community has devoted substantial effort to developing linear regression models for Interval-valued data, leading to a rich and extensive body of literature on this topic. The existing approaches primarily differ in how the regression model is formulated, the definitions adopted for linear combinations of Interval-valued variables, and the strategies employed to handle the non-negativity constraints of the interval ranges. Further distinctions arise in the depth and scope of the inferential analyses conducted, as well as in the proposed extensions to non-linear regression frameworks. A comprehensive overview of the principal contributions in this area can be found in de Carvalho et al. (2021) and the references therein. Additionally, efforts have been directed towards enhancing the robustness of classical linear regression estimators for interval data, addressing the sensitivity of traditional methods to outliers and model deviations; see Fagundes et al. (2013); Lima Neto and de Carvalho (2018) for an in-depth discussion and further references.

In this work, we propose a linear regression model in which both the response variable and the explanatory variables are intervals. Apart from an error term, the response variable is expressed as a linear combination of the explanatory variables using Moore's algebraic structure (see Girão Serrão et al. (2023) for further details). This formulation provides a clear structural interpretation of the regression coefficients, explicitly and jointly modeling the center and range of the response variable. To estimate the regression coefficients, we use the L_2 Wasserstein distance, and the regression model linking macrodata and microdata proposed in Oliveira et al. (2021), which allows for the explicit derivation of ordinary least squares estimators.

As in conventional least squares estimation, the derived estimators remain sensitive to outliers. To mitigate this limitation, we introduce robust M-estimators tailored for Interval-valued regression. Their performance is benchmarked against some of the existing approaches, including those proposed by Fagundes et al. (2013); Lima Neto and de Carvalho (2018).

Acknowledgements
We thank FCT - Fundação para a Ciência e Tecnologia, Portugal, through the project UIDB/04621/2020, with DOI: 10.54499/UIDB/04621/2020.

References

de Carvalho, F.A.T., Lima Neto, E.A., da N. Rosendo, U. (2021). Interval joint robust regression method. *Neurocomputing 465*, 265–288.

Fagundes, R.A.A., de Souza, R.M.C.R., and Cysneiros, F.J.A. (2013). Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence 26*(1), 564–573.

Girão Serrão, R., Oliveira, M.R., Oliveira, L. (2023). Theoretical derivation of interval principal component analysis. *Information Sciences 621*, 227–247.

Lima Neto, E.A., de Carvalho, F.A.T. (2018). An exponential-type kernel robust regression model for Interval-valued variables. *Information Sciences 454-455*, 419–442.

Oliveira, M.R., Azeitona, M., Pacheco, A., Valadas, R. (2021). Association measures for interval variables. *Advances in Data Analysis and Classification 15*, 1–30.

Integrating Centre and Range Methods in Symbolic Regression Trees

Priscilla Angulo[1,⋆], **Oldemar Rodríguez**[1]

1. CIMPA, School of Mathematics, University of Costa Rica
⋆Contact author: priscilla.angulo@ucr.ac.cr

Keywords: Symbolic data analysis, Interval-valued data, Regression trees, Centre and range method, Model optimization.

In symbolic data analysis, regression trees have been employed to model complex data structures, including variables represented as intervals. However, the use of the *centre and range* method as a representation strategy for Interval-valued variables has not yet been explored within the framework of symbolic regression trees. This gap presents an opportunity to extend current methodologies and evaluate whether this representation enhances predictive performance or model interpretability.

In this paper, we analyze the synthetic dataset `cardiologicalv2`, available in R. Symbolic regression trees are constructed using the `RSDA` package, which requires careful tuning of the `minsplit` and `maxdepth` parameters to optimize predictive accuracy. Various parameter combinations are evaluated by constructing a regression tree for each and computing the difference in root mean squared error (RMSE) between training and testing sets. The combination yielding the lowest prediction error is selected as optimal.

As an alternative, the centre and range components of the symbolic table are extracted, and two separate regression models are built using the `rpart` function—one for the centres and another for the ranges. Interval predictions are then reconstructed as $[\text{centre} - \text{range},\ \text{centre} + \text{range}]$, following the methodology introduced by Lima Neto and de Carvalho (Lima Neto and de Carvalho (2008)), who proposed this approach in the context of linear regression for Interval-valued symbolic data, reporting promising results on both synthetic and real datasets.

To compare the effectiveness of both approaches, RMSE is used as the evaluation metric, with the goal of minimizing predictive error.

The results reveal that the centre and range representation does not outperform the traditional symbolic regression tree. On the contrary, interval reconstructions yielded substantially higher RMSE values (lower bound RMSE = 19.14, upper bound RMSE = 11.74) compared to those obtained directly from symbolic trees (lower bound RMSE = 0.18, upper bound RMSE = 1.41). These findings suggest that decomposing the interval into separate components may lead to a significant loss of structural information, at least in the context of symbolic regression trees.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Ltd.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. CRC Press.

Dias, S. and Brito, P. (2022). Fundamental concepts about distributional data. In *Analysis of Distributional Data*.

Haddad, R. (2016). *Apprentissage supervisé de données symboliques et l'adaptation aux données massives et distribuées*. PhD thesis, Université Paris-Dauphine.

Lima Neto, E. and de Carvalho, F. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52(3), 1500–1515.

Mballo, C. and Diday, E. (2005). Decision trees on interval valued variables. *The Electronic Journal of Symbolic Data Analysis*, 3(1), 8–18. LISE-CEREMADE, Université Paris Dauphine.

Rodriguez, O. (2023). RSDA: R to Symbolic Data Analysis. R package version 3.2.1. https://CRAN.R-project.org/package=RSDA

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Zou, W. (2021). *Classification and Regression Trees for Symbolic Data*. PhD thesis, University of Georgia.

Regression Models with Sentiment Analysis Integration for Interval-valued Data

Daniela Sevilla[1], **Oldemar Rodríguez**[1]

1. CIMPA, School of Mathematics, University of Costa Rica
*Contact author: danisemo@hotmail.es

Keywords: Symbolic models, Regression models for interval-type data, Sentiment analysis, Tokenization, Natural language processing (NLP).

This paper introduces regression models that integrate sentiment analysis into Interval-valued data, aiming to improve predictive accuracy in contexts where data variability plays a critical role. The proposed methodology begins by transforming textual information into numerical representations, using methods based on predefined emotion dictionaries, commonly referred to as lexicons.

The process starts by creating a structured collection of documents, known as a *corpus* Feinerer et al. (2008), built using the `tm` package in *R* Feinerer and Hornik (2019). After constructing the corpus, the analysis applies a series of preprocessing steps, including the removal of punctuation, numbers, stopwords, and redundant whitespace, followed by lemmatization Addiga and Bagui (2022). It then constructs a term-document matrix $M \in \mathbb{N}^{m\times n}$, where each row represents a document and each column corresponds to a unique term. Each entry in the matrix represents the frequency with which the corresponding term appears in the text collection Feinerer et al. (2008).

Using the `textdata` package Silge and Robinson (2022) in *R*, the methodology retrieves the Afinn and NRC lexicons to generate a unified dictionary $\mathcal{L}$. It then divides this dictionary into two subsets based on term polarity: one containing negatively connoted terms, and the other, positively connoted terms. These subsets are used to filter the term-document matrix M, yielding two new matrices: M^-, which includes only negatively connoted terms, and M^+, which includes only positively connoted ones.

The analysis enhances semantic sensitivity by incorporating contextual modifiers such as negators, intensifiers, and attenuators. These modifiers alter both the polarity and emotional intensity of terms, so the model adjusts term frequencies accordingly. It uses the *hash_valence_shifters* lexicon from the `lexicon` package Rinker (2023) in *R* to identify such modifiers.

If a term is preceded by a negator, the analysis inverts its polarity and moves it to the opposite matrix. It also adjusts frequency values: it adds 0.2 when an intensifier is present, subtracts 0.2 when the term is attenuated and leaves the value unchanged if no modifier applies. This procedure produces adjusted matrices $M^{+\text{adjusted}}$ and $M^{-\text{adjusted}}$, which more accurately represent the emotional content of the text.

From these matrices, the analysis constructs an interval for each document d_i:

$$I_i = \left[-\sum_{j=1}^{p} M_{ij}^{-\text{adjusted}}, \quad \sum_{j=1}^{q} M_{ij}^{+\text{adjusted}} \right]$$

where p and q denote the number of positive and negative terms, respectively. The lower bound captures the accumulated intensity of negative connotations (expressed with a negative sign to

reflect emotional opposition), while the upper bound reflects the total of adjusted positive connotations. The resulting interval becomes a new symbolic variable in the dataset.

To evaluate the impact of sentiment analysis, the study compares model performance with and without the inclusion of this symbolic variable across several regression methods. These include the Center (CR) and Center and Range (CRM) approaches for linear regression Lima and De Carvalho (2010), along with regularized variants using Ridge and Lasso penalties Rodríguez (2018). The evaluation also considers non-parametric techniques such as K-Nearest Neighbors (KNN) and neural networks Lima-Neto and De Carvalho (2017) and Rodríguez (2018).

Performance is assessed using evaluation metrics proposed by Lima and De Carvalho (2010), which include the root mean square error (RMSE) for both interval bounds, the correlation coefficient for each bound, and the coefficient of determination R^2.

The implementation uses the `RSDA` package Rodríguez et al. (2023), which provides functions for fitting regression models to Interval-valued data and calculating the evaluation metrics. The results demonstrate improved predictive performance when sentiment analysis is integrated. The validation uses the *Hotel Reviews Dataset Enriched*, available on Kaggle Calislar (n.f.).

References

Addiga, A., Bagui, S. (2022). Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency. *Journal of Computer and Communications 10*, 117–128. https://doi.org/10.4236/jcc.2022.108008

Calislar, Y. (n.f.). Hotel Reviews Dataset Enriched. https://www.kaggle.com/datasets/ycalisar/hotel-reviews-dataset-enriched.

Feinerer, I., Hornik, K., Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software 25*, 1–54. https://api.semanticscholar.org/CorpusID:51738608.

Feinerer, I., Hornik, K. (2019). Text Mining Package 'tm'. https://cran.r-project.org/web/packages/tm/index.html.

Lima, E., De Carvalho, F. (2010). Constrained linear regression models for symbolic Interval-valued variables. *Computational Statistics & Data Analysis 54*(2), 333–347. https://doi.org/10.1016/j.csda.2009.08.010.

Lima-Neto, E. A., De Carvalho, F. A. T. (2017). Nonlinear regression applied to Interval-valued data. *Pattern Analysis and Applications 20*, 809–824. https://doi.org/10.1007/s10044-016-0538-y.

Rinker, T. (2023). lexicon: Dictionaries, Lookup Tables, and Lexicons for Text Analysis (Versión 1.2.1) [Paquete de R]. https://CRAN.R-project.org/package=lexicon.

Rodríguez, O. (2018). Shrinkage linear regression for symbolic Interval-valued variables. *Journal MODULAD 45*, 19–38.

Rodríguez, O., Chacon, J. E., Aguero, C., Arce, J. (2023). Package 'RSDA': R to Symbolic Data Analysis (Version 3.2.1). https://oldemarrodriguez.com/. Symbolic Data Analysis (SDA) implements various techniques of automatic classification and linear models for symbolic data.

Silge, J., & Robinson, D. (2022). textdata: Download and Load Text Datasets (Versión 0.4.4) [Paquete de R]. https://CRAN.R-project.org/package=textdata.

Supervised Learning

A New Discriminant Analysis Approach for Density-valued Data Classification

Francesca Condino[1*], **Paula Brito**[2]

1. Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Italy
2. Faculdade de Economia, Universidade do Porto & LIAAD - INESC TEC, Portugal
*Contact author: francesca.condino@unical.it

Keywords: Parametric distribution, Mixture, Kullback-Leibler divergence, Entropy.

In recent decades, the logic underlying traditional methods used to perform discriminant analysis for classifying a set of statistical units has been extended to more complex data structures. These structures involve data recorded, for example, as intervals, histograms or distributions, within to a symbolic data table. Symbolic Data Analysis is particularly useful for handling datasets containing more complex and structured information than conventional unit-variable data tables.

In this framework, discriminant analysis aims at establishing a decision rule that effectively separates different groups within a symbolic dataset. Some proposals for extending and adapting discriminant analysis to accommodate symbolic structures are already present in the literature (Ishibuchi et al., 1990; Duarte Silva and Brito, 2006, 2015; Dias et al., 2021), and often refer to linear discriminant methods. These methods rely on different types of symbolic representations and techniques, and consider specific issues such as measuring the distance among objects to manage the inherent nature and complexity of symbolic data. Indeed, one major challenge is defining appropriate distance measures to ensure the effectiveness of the discriminant process.

In this study, we aim at classifying symbolic objects described by parametric density functions into two predefined groups. To this end, we employ the Jensen-Shannon divergence as dissimilarity measure between objects. This is an entropy-based measure, also related to the Kullback-Leibler divergence, and has already been used in the context of density function clustering (Condino, 2009, 2023) due to its advantageous properties. Specifically, it can be shown that adopting this measure allows us to obtain the barycentre of each group as a mixture of densities describing the units within that group. Therefore, the barycentre is still a density function, so that each centre belongs to the space of description of the considered symbolic objects. In this context, it is possible to verify that the total divergence, i.e. the divergence of all considered objects, can be decomposed in two components, one relating to the dissimilarities within each group and the other reflecting the dissimilarities between groups, according to Huygens' theorem. Based on this evidence, it is possible to derive a classification rule to assign each statistical unit to one of the two groups, ensuring that each symbolic object is allocated to the group in such a way that the total Jensen-Shannon divergence is minimized.

An application to real data is performed. The data pertain to air time and departure delays of the airline companies operating in NY airports in 2013, classified as Main and Regional carriers. These data have already been used in the context of linear discriminant analysis of distributional data (Dias et al., 2021), by considering interval and histogram-valued variables. Here, a symbolic data table is constructed where each unit is the airline/month and each descriptor is a parametric density function, obtained by fitting a specific parametric model on monthly data. In particular, the Generalized Extreme Value distribution is considered to model Departure Delay, while the model proposed by Domma and Condino (2014) is used to describe Air Time distribution. This latter

model is particularly useful in this case, as it is well-suited for describing data characterized by multiple modal points, as observed here. For each of the two groups, the barycentre is obtained as the mixture of the densities within the group, and the classification into the two groups is performed, according to the specified rule and by using numerical integration method to compute the dissimilarity values. The obtained results show a good performance of the proposed method in discriminating between the two types of carriers, achieving an accuracy above 87% for each descriptor.

References

Condino, F. (2009). *La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico per dati descritti da distribuzioni multivariate.* Ph.D. thesis, University of Naples "Federico II".

Condino, F. (2023). Share density-based clustering of income data. *Statistical Analysis and Data Mining: The ASA Data Science Journal 16 (4)*, 336–347.

Dias, S., Brito, P., Amaral P. (2021). Discriminant analysis of distributional data via fractional programming. *European Journal of Operational Research 294 (1)*, 206–218.

Domma, F., Condino, F. (2014). A new class of distribution functions for lifetime data. *Reliability Engineering & System Safety 129*, 36–45.

Ishibuchi, H., Tanaka, H., Kukuoka, N. (1990). Discriminant Analysis of multi-dimensional Interval Data and its application to chemical sensing. *International Journal of General Systems 16(4)*, 311-329.

Duarte Silva, A.P., Brito, P. (2006). Linear Discriminant Analysis for Interval Data. *Computational Statistics 21(2)*, 289–308.

Duarte Silva, A.P., Brito, P. (2015). Discriminant Analysis of Interval Data: An Assessment of Parametric and Distance-Based Approaches. *Journal of Classification 32(3)*, 516-541.

Stacked Logistic Regression for Interval Data Classification

Rafaella L. S. do Nascimento[1*], Renata M. C. R. de Souza[1], Francisco José A. Cysneiros[2]

1. Centro de Informática, Universidade Federal de Pernambuco, Brazil
2. Departamento de Estatística, Universidade Federal de Pernambuco, Brazil
*Contact author: rlsn@cin.ufpe.br

Keywords: Interval-valued data, Ensemble model, Logistic regression.

Symbolic Data Analysis (SDA) extends traditional data analysis and mining techniques to handle more complex data types, such as Interval-valued data, representing variables using lower and upper bounds (Billard and Diday, 2012). Interval data are particularly useful in economics, meteorology, and social sciences, where observed phenomena naturally fluctuate within well-defined ranges. Interval-based classification models extend traditional algorithms, such as logistic regression, decision trees, and neural networks, to handle the added complexity of Interval-valued inputs. As presented in Souza, Queiroz and Cysneiros (2011), logistic regression classifiers in SDA for the interval data often use the center and the interval bounds, with predictions typically based on the average of the lower and upper bound estimates.

This study explores a new logistic regression classification rule for Interval-valued data. The proposed methodology transforms the lower and upper bounds into separate predictor variables, which serve as input for multiple base models. The predictions of these models are then combined into a metamodel that uses logistic regression to capture interactions between the components of the interval and improve predictive accuracy (Breiman, 1996). Although averaging the bounds may overlook information about data dispersion, the stacked model learns how each bound contributes to the prediction separately, allowing the metamodel to integrate the information more effectively.

Definition 1 Let K be the number of pattern classes labeled as $1, \ldots, K$. A symbolic training set $\mathcal{X} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ consists of N instances, where each $x^{(i)}$ is a vector of p symbolic interval variables $X_j^{(i)} = [a_{ij}, b_{ij}]$ and $y^{(i)}$ is a categorical response variable taking values in $\{1, \ldots, K\}$. Each interval variable X_j belongs to the set $\mathcal{I} = \{[a, b] \in \mathbb{R}^2 : a \leq b\}$. The categorical response Y is represented by K binary indicator variables Y_k, such that $Y_k = 1$ if $Y = k$, and $Y_k = 0$ otherwise. Each Y_k follows a Bernoulli distribution with parameter $p_k = \Pr(Y = k \mid x)$. A set of K linear functions $f_k(x)$ is used within a multiclass logistic regression framework to estimate these posterior probabilities. This approach applies a one-vs-rest strategy.

Definition 2 Consider a stacking-based model where the lower and upper bounds of the interval data are used as features for a meta-model based on logistic regression. Let the interval data for each pattern i be represented by two vectors of p covariates: $\mathbf{x}_L = (x_{L1}, \ldots, x_{Lp})$ and $\mathbf{x}_U = (x_{U1}, \ldots, x_{Up})$, where $x_{Lj} = a_j$ and $x_{Uj} = b_j$ are the lower and upper bounds, respectively.

The first level of the stacked logistic model trains two base classifiers (logistic regression) separately, using $\mathbf{x}_L$ and $\mathbf{x}_U$ as input. Each base classifier provides a probability prediction for each class k, denoted as $P_k(\mathbf{x}_L)$ and $P_k(\mathbf{x}_U)$. These predictions form a vector of characteristics $\mathbf{z} = (P_1(\mathbf{x}_L), P_2(\mathbf{x}_L), \ldots, P_K(\mathbf{x}_L), P_1(\mathbf{x}_U), P_2(\mathbf{x}_U), \ldots, P_K(\mathbf{x}_U))$, input to a metamodel.

The metamodel is based on logistic regression and is trained to predict the final class $\hat{k}$ for a given input $\mathbf{x}$. The logistic regression model uses the feature vector $\mathbf{z}$ as input and predicts the class probabilities $P_k(\mathbf{z})$ for each class k. The predicted class $\hat{k}$ is the one with the highest probability $P_k(\mathbf{z})$ among all classes k, given the feature vector $\mathbf{z}$. Figure 1 shows the classification steps.

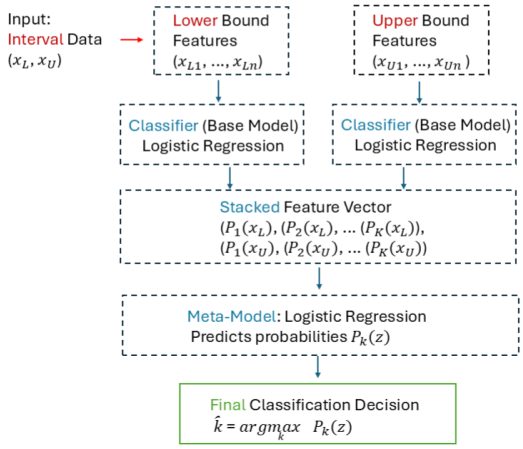


Figure 1: Steps of classification rule based on Stacked Logistic Regression.

Example: Two seed datasets were constructed, each containing 1200 points distributed into three unbalanced classes: class 1 $n = 600$; class 2 $n = 400$; class 3 $n = 300$. The datasets were generated using bivariate normal distributions with the following characteristics in Table 1:

Table 1: Parameters of Seed Datasets I and II

Class	**Seed Dataset I: Well-Separated**	**Seed Dataset II: Overlapping**
1	$\boldsymbol{\mu} = (15, 5)^\top, \sigma_1^2 = 64, \sigma_2^2 = 9, \sigma_{12} = 0$	$\boldsymbol{\mu} = (17, 5)^\top, \sigma_1^2 = 81, \sigma_2^2 = 25, \sigma_{12} = 0$
2	$\boldsymbol{\mu} = (30, 10)^\top, \sigma_1^2 = 25, \sigma_2^2 = 36, \sigma_{12} = 0$	$\boldsymbol{\mu} = (25, 15)^\top, \sigma_1^2 = 16, \sigma_2^2 = 81, \sigma_{12} = 0$
3	$\boldsymbol{\mu} = (5, 10)^\top, \sigma_1^2 = 25, \sigma_2^2 = 9, \sigma_{12} = 0$	$\boldsymbol{\mu} = (15, 10)^\top, \sigma_1^2 = 16, \sigma_2^2 = 16, \sigma_{12} = 0$

Bivariate data classes were generated from seed vectors $(s_1, s_2)^\top$, where each class size n was randomly drawn from a uniform distribution $U[20, 60]$. The individual units within each class were sampled from a bivariate normal distribution with independent components. The data values for each component were also simulated using uniform distributions with five different ranges: $U[1, 10]$, $U[1, 20]$, $U[1, 30]$, $U[1, 40]$, and $U[1, 50]$.

The analysis followed a Monte Carlo approach with 500 replications per dataset. For each replication, 75% of the data were randomly selected for training and 25% for testing. The classification error was computed per class on the test set, and the final error rate was obtained by averaging these values. Table 2 shows the results. As the interval range increases (from [1-10] to [1-50]), the classification error of both models slightly increases; however, the stacked logistic model outperforms the average model across all configurations.

Table 2: Mean classification error of models in Criteria 1 for both scenarios

	Dataset I – Well-Separated Classes					**Dataset II – Overlapping Classes**				
Model	[1–10]	[1–20]	[1–30]	[1–40]	[1–50]	[1–10]	[1–20]	[1–30]	[1–40]	[1–50]
average	0.175	0.175	0.179	0.180	0.183	0.340	0.342	0.344	0.344	0.345
stacked	**0.169**	**0.171**	**0.173**	**0.175**	**0.177**	**0.335**	**0.337**	**0.339**	**0.339**	**0.340**

References

Billard, L., Diday, E. (2012). *Symbolic data analysis: Conceptual statistics and data mining.* John Wiley & Sons.

Souza, R. M. C. R., Queiroz, D. C. F., Cysneiros, F. J. A. (2011). Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications, 14(3)*, 273-282.

Breiman, L. (1996). Stacked regressions. *Machine learning, 24*, 49-64.

Applications of SDA

LIMOS - LightGBM Interval Merton's One-period-portfolio Selection

Liang-Ching Lin[1,⋆]

1. Department of Statistics, National Cheng Kung University, Tainan, Taiwan
⋆Contact author: lclin@ncku.edu.tw

Keywords: LightGBM, Merton's portfolio problem, Symbolic data analysis.

The modern portfolio theory can assist us in allocating wealth to risky and risk-free assets reasonably by using some statistical methods. In this study, we will focus on evolving Merton's portfolio problem proposed by Merton (1969). By maximizing the expected utility function of the portfolio value processes, Merton (1969) obtained the optimal portfolio weight but required the estimation of the mean and variance of the log returns. Instead of the conventional parameter estimations based on only the closing prices, referring to Lin and Sun (2019), we include the opening, high, low, and closing prices to enlarge the database as much as possible to make the parameter estimations much more accurate. Furthermore, the covariances are estimated using the bivariate symbolic Interval-valued variables under a copula function as shown in Lin, Guo and Lee (2023). However, we found that the estimation of the mean of the log returns is relatively inaccurate such that we may have the incorrect transaction direction. That is, we may decide to buy the stock but it falls or vice versa and then we lose the money. In order to solve this problem, we use the LightGBM to predict the transaction directions, in which, the stock prices and many statistics related to Interval-valued variables are included in the features. In real data analysis, we demonstrate the usefulness of combining the aforementioned methods by showing the portfolio profits of selecting 10 stocks in 2018 and 2019. The results particularly show the superiority of the proposed strategy over the conventional method: the profits are almost positive and have around 60% to 94% annually.
(This work is joint with Hao-Chien Huang and Sz-Wei Charng).

References

Liang-Ching, Lin, Meihui, Guo, Sangyeol, Lee. (2023). Monitoring photochemical pollutants based on symbolic Interval-valued data analysis *Advances in Data Analysis and Classification 17*, 897-926.

Liang-Ching Lin and Li-Hsien Sun. (2019). Modeling Financial Interval Time Series *PLoS ONE 14(2)*, e0211709.

Robert C. Merton. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The review of Economics and Statistics*, 247-257.

Symbolic Data Analysis Approach to Identify adolescent Profiles Based on Momentary Self-assessments and the Use of Internet Applications

Jasminka Dobša[1,*], Simona Korenjak-Černe[2,3], Miranda Novak[4], Maja Buhin Pandur[1]

1. University of Zagreb Faculty of Organization and Informatics, Zagreb, Croatia
2. University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia
3. Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
4. University of Zagreb Faculty of Education and Rehabilitation Sciences, Department of Behavioural Disorders, Zagreb, Croatia

*Contact author: jasminka.dobsa@foi.hr

Keywords: Adolescent, Ecological momentary assessment, Passive data, Effortless Assessment Research System (EARS), Symbolic data.

The aim of our research is to develop a metodology for exploratory data analysis of symbolic data to test the positive youth development framework using traditional and digital mobile assessment. In the presentation we intend to discuss some conceptual ideas and present some preliminary results using a real data set.

Data: One hundred and thirty Croatian high-school students reported on the quality of their close friendships and their affect seven times a day for one week (i.e. 49 assessments). More than half of them (58 %) were female, 40 % were male and 2 % preferred not to state their gender. Their average age was 15.91 years (SD = .314). While 60 % of them were attending secondary school preparing them for higher education, 40 % of them were in vocational training. The schedule for the assessments was semi-random, meaning that participants were asked the questions at random intervals in two-hour blocks of time between 7am and 9pm.

Data collection: The study was conducted using the Effortless Assessment Research System (EARS) application from Ksana Health, University of Oregon (Lind et al. (2018)), which allows for a combination of ecological momentary assessment questions (i.e., about participants' experiences and behaviors in the current moment in time) and passive mobile data collection. The collected passive data contains information on the use of a total of 927 mobile applications used by the responders during the observation period. These applications were categorized into 16 groups, which were formed semi-automatically using generative AI (ChatGPT, Google Bard): Books and Reading, Communication, Device Management, Education and Learning, Entertainment, Finance and Banking, Games, Health and Fitness, Multimedia, Music and Audio, News, Online Shopping and Services, Social Media, Tools and Productivity, Travel, and Other.

The obtained data set consists of three groups of variables: 1) survey data (gender, school success, risk level for depression and anxiety, and the like), 2) assessment questions collected from responders using the EARS application about the quality of their friendship relationships, mood, skills, self-perception (awareness), and similar ecological momentary assessment data, and 3) variables about the total amount of time spent with specific groups of mobile applications (passive data). The data have already been analysed using factor analysis to analyse the quality of friendship and well-being in adolescence based on daily active data (Šutić et al. (2025)).

The aim of this study is to identify profiles of adolescents based on ecological momentary as-

sessment and passive data using clustering methods of symbolic data analysis (Billard and Diday (2019), Brito and Dias (2022)). Since each responder answered the ecological momentary assessment questions several times, these data can be considered as symbolic data (Bock and Diday (2000), Brito (2014), Diday (2016)). For this purpose, we will include as the subjects (observed units) responders with at least 10 assessments (recordings) and describe them with symbolic data. Considering relations between 25 measured variables of the ecological momentary assessment data, six composed variables are defined. These variables can be expressed either as weighted mean values or as a distribution of data ranging from 1 to 7, where the values are weighted according to the calculated composite frequencies. Such distributional data fit to the framework of symbolic data and enable to preserve intrinsic variability of the answers. For them, several dissimilarity measured have already been proposed (Brito and Dias (2022)). The variables related to the passive data, on the other hand, can be considered as total applets usage or interval-valued data. To combine both types of variables, passive data can be used as external variables, or a composite distance measure should be found that combines both types, ecological momentary and passive data. All these options allow us to use classical clustering approaches or symbolic clustering approaches. In practical application, however, the question arises as to which approach provides better results. Since the evaluation in this case is not straightforward, we will discuss both options.

Acknowledgements
This work is part of the project "Testing the 5C framework of positive youth development: traditional and digital mobile assessment - P.R.O.T.E.C.T." funded by the Croatian Science Foundation (UIP-2020-02-2852) and was also partially supported by grant P1-0294 from ARIS, Slovenia.

References

Billard, L, Diday, E. (2019). Clustering Methodology for Symbolic Data. *Wiley Series in Computational Statistics* John Wiley & Sons, Ltd.

Bock, L, Diday, E. (Eds.) (2000). Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg.

Brito, P. (2014) Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(4)*, 281–295.

Brito, P., Dias, S. (Eds.) (2022). Analysis of Distributional Data. Chapman and Hall CRC.

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

Lind, M. N., Byrne, M. L., Wicks, G., Smidt, A. M., Allen, N. B. (2018). The Effortless Assessment of Risk States (EARS) Tool: An Interpersonal Approach to Mobile Sensing. *JMIR Ment Health 5(3)*:e10334. DOI: 10.2196/10334. PMID: 30154072; PMCID: PMC6134227.

Šutić, L., van Roekel, E., Novak, M. (2025). Quality of friendships and well-being in adolescence: daily life study. *International Journal of Adolescence and Youth 30(1)*, 2467112. DOI: 10.1080/02673843.2025.2467112.

Index of Authors

Acknowledgements

The Organizing Committee expresses its sincere appreciation to all participants of the workshop for their active engagement, insightful contributions, and commitment to the exchange of knowledge throughout the event.

We extend our special thanks to the distinguished panelists of the discussion *"New Horizons in Official Statistics: Techniques, Tools, and Challenges"*:

- **Ivana Levačić**, Croatian Bureau of Statistics, Sector for Statistical Methodologies, Quality and User Relations
- **Jure Dubravčić**, Croatian Bureau of Statistics, Sector for Information Technologies
- **Boris Berenček**, Teched Consulting Services Ltd.
- **Prof. Paula Brito**, University of Porto, Portugal
- **Prof. Oldemar Rodriguez**, University of Costa Rica

whose expertise and thoughtful perspectives greatly enriched the panel and fostered meaningful discussion on the future of official statistics.

We are also grateful to the presenters of the tutorial and software sessions for sharing their methodological developments and practical tools:

- **Prof. Paula Brito**, University of Porto, Portugal – *Tutorial: Symbolic Data Analysis – Why, How, What for?*
- **Prof. Oldemar Rodriguez**, University of Costa Rica – *Software Presentation: The RSDA Package*
- **Prof. Pedro Duarte Silva**, Universidade Católica Portuguesa, Portugal – *Software Presentation: MAINT.Data*
- **Prof. Antonio Irpino**, University of Campania "L. Vanvitelli", Italy – *Software Presentation: HistDAWass*

To all contributors and attendees, thank you for helping make this workshop a productive, collaborative, and inspiring event.

Notes:


foi