

Javier Arroyo  
Carlos Maté  
Paula Brito  
Monique Noirhomme-fraiture (Eds.)

# 3<sup>rd</sup> Workshop in Symbolic Data Analysis

Madrid, 7-9 November 2012

Book of Abstracts



---

# 3<sup>rd</sup> Workshop in Symbolic Data Analysis

---



Universidad Complutense  
de Madrid



J. Arroyo  
C. Maté  
P. Brito  
M. Noirhomme-Fraiture (eds.)

November, 2012

Documento maquetado con T<sub>E</sub>X<sup>I</sup>S v.1.0+.

# 3<sup>rd</sup> Workshop in Symbolic Data Analysis

Book of Abstracts

**November, 2012**

ISBN 978-84-695-6575-9

Depósito legal: M-38333-2012

# Contents

Preface	v
<b>I Clustering I</b>	
November 7, 11:30 - 13:00	<b>1</b>
<b>II Visualization &amp; Multidimensional Scaling</b>	
November 7, 14:30 - 16:30	<b>9</b>
<b>III Applications I</b>	
November 7, 17:00 - 18:30	<b>19</b>
<b>IV Regression</b>	
November 8, 9:00 - 11:00	<b>27</b>
<b>V Dimensionality Reduction</b>	
November 8, 11:30 - 13:00	<b>37</b>
<b>VI Time Series</b>	

November 8, 14:30 - 16:30	45
VII Software	
November 8, 17:00 - 18:00	55
VIII Applications II	
November 9, 9:00 - 11:00	61
IX Clustering II	
November 9, 11:30 - 13:00	71

# Discriminant Analysis of Interval Data: Parametric Versus Distance-Based Approaches

A. Pedro Duarte Silva<sup>1,\*</sup>, Paula Brito<sup>2</sup>

1. Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal

2. Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal

\*Contact author: psilva@porto.ucp.pt

**Keywords:** Discriminant analysis, Interval data, Parametric modelling of interval data

In this paper, we are interested in the analysis of interval data, i.e., where elements are characterized by variables whose values are intervals on  $\mathbb{R}$ , and investigate and compare different methods or discriminant analysis of such data.

Distance-based approaches to linear discriminant analysis of interval data are discussed in Duarte Silva & Brito (2006). These approaches lead to representations in the discriminant space in the form of intervals or single points, from which distance-based allocation rules are derived. In Brito & Duarte Silva (2012) a parametric modelling for interval data, assuming multivariate Normal or Skew-Normal distributions for the Midpoints and Log-Ranges of the interval variables, is proposed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, represented by five different possible configurations. This approach is implemented in an R package, MAINT.DATA Duarte Silva and Brito (2011) (available at the CRAN repository), which includes several tools for modelling and analysing interval data. In particular MAINT.DATA introduces a data class for representing interval data and provides methods and functions for parameter estimation, statistical tests for the different covariance configurations, and parametric discriminant analysis of interval data.

Discriminant analysis of interval data has been investigated by other authors in different contexts. Ishibuchi, Tanaka and Noriko Fukuoka (see Ishibuchi, Tanaka & Fukuoka (1990)) address discriminant analysis of interval data determining interval representations in a discriminant space using a mathematical programming formulation. Approaches of discriminant analysis of interval data based on imprecise probability theory may be found in Nivlet, Fournier & Royer (2001) and Utkin & Coolen (2011). In Lauro, Verde & Palumbo (2000), a generalization of classical Factorial Discriminant Analysis to symbolic data is proposed. This method is based on a numerical analysis of the transformed symbolic data, followed by a symbolic interpretation of the results; it allows considering quantitative, qualitative nominal or modal variables; classification rules are then based on proximities in the factorial plane (see also Lauro, Verde & Irpino (2008)).

This paper evaluates the relative performance of different classification rules for interval data. It compares the distance-based classification rules considered in Duarte Silva & Brito (2006), the parametric classification rules derived from the models discussed in Brito & Duarte Silva (2012), and rules proposed by other authors.

Preliminary results show that parametric approaches generally outperform other approaches, and that restricted configurations of the variance-covariance matrix which take into account the particular nature of interval data lead to parsimonious rules, which can be quite effective in reducing expected error rates.

## Acknowledgments

This work is partly funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação

## Dimensionality Reduction

para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

### References

- Brito, P. and Duarte Silva, A.P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, 39 (1), 3-20.
- Duarte Silva, A.P. and Brito, P. (2006). Linear discriminant analysis for interval data. *Computational Statistics*, 21 (2), 289-308.
- Duarte Silva, A.P. and Brito, P. (2011). **MAINTData: Model and Analze Interval Data**. R package version 0.2, URL <http://cran.r-project.org/package=MAINT.Data>.
- Ishibuchi, H., Tanaka, H. and Fukuoka, N. (1990). Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *International Journal of General Systems*, 16 (4), 311-329.
- Lauro, N.C., Verde, R. and Palumbo, F. (2000). Factorial discriminant analysis on symbolic objects. In: Bock, H.-H. and Diday, E. (Eds.), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg, 212-233.
- Lauro, N.C., Verde, R. and Irpino, A. (2008). Factorial discriminant analysis, in In: Diday, E. and Noirhomme-Fraiture, M. (Eds.), *Symbolic Data Analysis and the Sodas Software*, Wiley, Chichester, 341-358.
- Nivlet, P., Fournier, F. and Royer, J.J. (2001). Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. In: *ISIPTA'01*, 284-292 .
- Utkin, L.V. and Coolen, F.P.A. (2011). Interval-valued regression and classification models in the framework of machine learning. In: *7th International Symposium on Imprecise Probability: Theories and Applications*. Innsbruck, Austria.