6th International Workshop

Symbolic Data Analysis

2017

ABSTRACTS and PROGRAM

June 12 - 14, 2017
Ljubljana, Slovenia

http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:lj17

6th International Workshop

Symbolic Data Analysis

2017

ABSTRACTS and PROGRAM

June 12 - 14, 2017

Ljubljana, Slovenia

http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:lj17

Organized by

Statistical Society of Slovenia,
University of Ljubljana – Faculty of Economics,
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia

Supported by

Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia

Statistical Society of Slovenia

SURS – Statistical Office of Republic of Slovenia

University of Ljubljana – Faculty of Economics

Dragon Bridge with the Castle in the back (Ljubljana)

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

311(082)

INTERNATIONAL Workshop Symbolic Data Analysis (6 ; 2017; Ljubljana)
Abstracts and program / June 12 - 14, 2017, Ljubljana, Slovenia organized by Institute of Mathematics, Physics and Mechanics, Ljubljana [and] Statistical Society of Slovenia [and] University of Ljubljana, Faculty of Economics; [edited by Vladimir Batagelj, Simona Korenjak-Černe and Nataša Kejžar]. - Ljubljana : Statistical Society of Slovenia, 2017

ISBN 978-961-93547-8-0

1. Dodat. nasl. 2. Batagelj, Vladimir 3. Statistično društvo Slovenije 4. Inštitut za matematiko, fiziko in mehaniko (Ljubljana) 5. Ekonomska fakulteta (Ljubljana)
290302464

Scientific Program Committee

Javier Arroyo (Spain)
Lynne Billard (USA)
Chun-Houh Chen (Taiwan)
Edwin Diday (France)
Rosanna Verde (Italy)
Vladimir Batagelj (Slovenia)
Paula Brito (Portugal)
Francisco de Carvalho (Brasil)
Monique Noirhomme-Fraiture (Belgium)

Organizing Committee

Vladimir Batagelj (Chair)
Nataša Kejžar
Simona Korenjak-Černe
Jerneja Čuk

Published by: Statistical Society of Slovenia
Litostrojska cesta 54
1000 Ljubljana, Slovenia
Edited by: Vladimir Batagelj, Simona Korenjak-Černe and Nataša Kejžar
Produced using: generbook R package
Circulation: 50

USEFUL INFORMATION

The workshop is held at

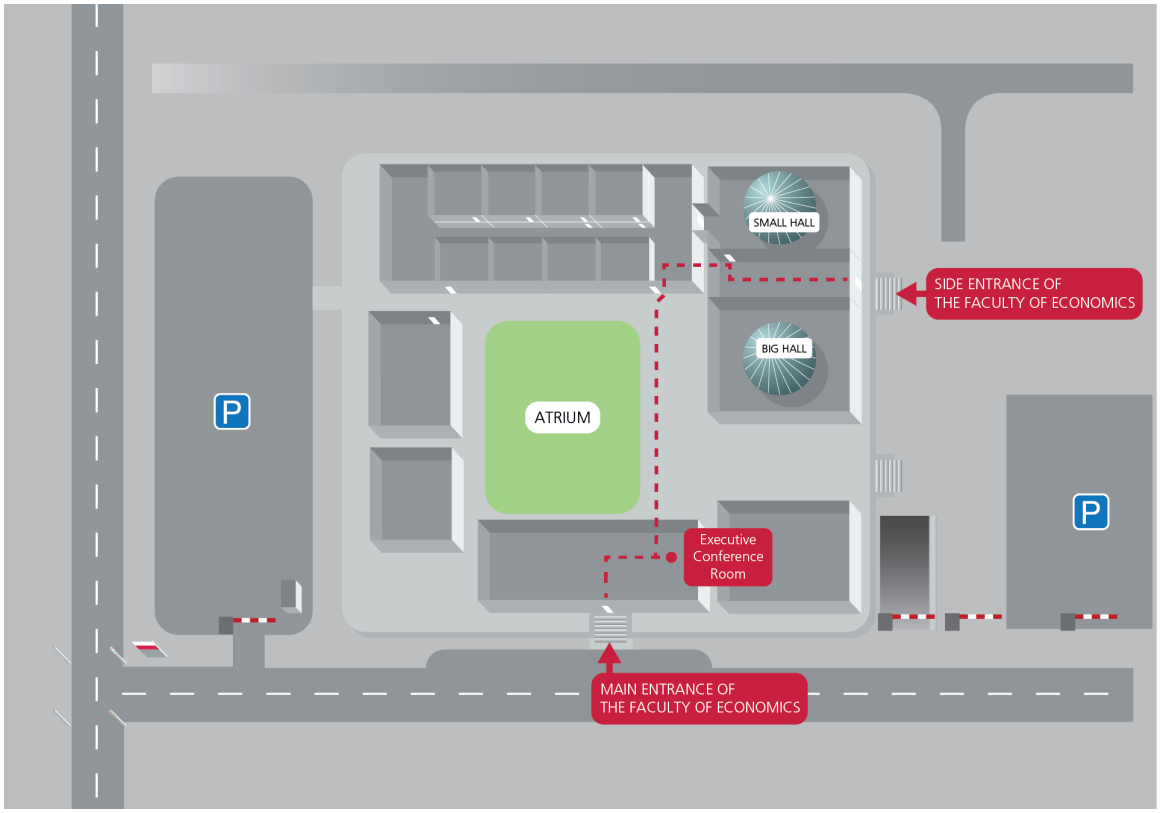
University of Ljubljana, Faculty of Economics
Kardeljeva ploščad 17
1000 Ljubljana
Closest City bus stations Mercator (lines 6, 8 and 11) and Kardeljeva ploščad (lines 13 and 20).

Lunches are in the same building as the workshop (Faculty of Economics).
Workshop Dinners are at the following two places

Monday, June 12	Tuesday, June, 13
Castle Restaurant Gostilna na Gradu Grajska planota 1 1000 Ljubljana	**Gostilna Čad** Cesta na Rožnik 18 1000 Ljubljana

Map of workshop venue at the Faculty of Economics (EF) is on the next page.

The workshop room is named Executive Conference Room. The refreshment breaks will be held in the hall in front of the Room.



PROGRAM

		workshop room at EF
Monday	11.00 – 11.30	Registration
	11.30 – 11.45	Opening of the Conference
	11.45 – 13.30	Lunch *Restaurant EF*
	13.30 – 15.00	General SDA
	15.00 – 15.30	Coffee Break
	15.30 – 17.30	Likelihood Functions
	19.30	Workshop Dinner *Castle Restaurant - Gostilna na gradu*
Tuesday	9.00 – 10.30	PCA, FA
	10.30 – 11.00	Coffee Break
	11.00 – 12.00	Visualization
	12.00 – 13.30	Lunch *Restaurant EF*
	13.30 – 14.30	Time Series
	14.30 – 15.00	Coffee Break
	15.00 – 17.00	Applications, Clustering
	19.30	Workshop Dinner *Gostilna Čad*
Wednesday	9.00 – 10.00	Software
	10.00 – 10.30	Coffee Break
	10.30 – 12.30	Applications, Networks
	12.30 – 14.00	Lunch *Restaurant EF*
	14.00 – 15.00	Round Table
	15.00	Closing

11.00–11.30 **Registration**

11.30–11.45 **Opening of the Conference** (EF)

11.45–13.30 **Lunch *Restaurant EF***

13.30–15.00 **General SDA** (EF) *Chair: Lynne Billard*

1. **Introduction to the Symbolic Data Analysis framework and application to post clustering for comparing and improving clustering methods by the "Symbolic Data Table" that they induce**
   Edwin Diday
2. **Intervals, histograms, quantiles, compositions... what on earth should I do?**
   Paula Brito
3. **Meta-analysis and SDA**
   Masahiro Mizuta

15.00–15.30 **Coffee Break**

15.30–17.30 **Likelihood Functions** (EF) *Chair: Paula Brito*

1. **Estimating a mixture of Dependent Dirichlet Distributions**
   Richard Emilion
2. **A general framework for constructing symbolic likelihood functions**
   Scott A Sisson, Boris Beranger and Huan (Jaslene) Lin
3. **Estimating species abundance using symbolic data meta-analysis**
   Huan (Jaslene) Lin, Julian Caley and Scott Sisson
4. **A composite likelihood based approach for max-stable processes using histogram-valued variables**
   Thomas Whitaker, Boris Beranger and Scott Sisson

19.30 **Workshop Dinner *Castle Restaurant - Gostilna na gradu***

9.00–10.30 **PCA, FA** (EF) *Chair: Richard Emilion*

1. **Symbolic covariance for interval data: Properties and interpretation**
   M. Rosário Oliveira, Margarida Vilela and António Pacheco
2. **Maximum likelihood factor analysis of interval data**
   Paula Cheira, Paula Brito and A. Pedro Duarte Silva
3. **Optimized symbolic principal component for interval-valued variables**
   Jorge Arce and Oldemar Rodríguez

10.30–11.00 **Coffee Break**

11.00–12.00 **Visualization** (EF) *Chair: Simona Korenjak-Černe*

1. **Matrix decompositions based on induced norms with application to interval data**
   Vartan Choulakian
2. **A visualization synthesis through symbolic data analysis of open ended questions**
   Mireille Gettler Summa, Myriam Touati and Sadika Rjiba

12.00–13.30 **Lunch *Restaurant EF***

13.30–14.30 **Time Series** (EF) *Chair: Scott Sisson*

1. **Data quality and hydrologic interval time series analysis**
   Carlo Drago, Silvia Di Francesco and Veronica Ciccone
2. **Forecasting wind speed distributions with kernel regression**
   Albert Meco and Javier Arroyo

14.30–15.00 **Coffee Break**

15.00–17.00 **Applications, Clustering** (EF) *Chair: Edwin Diday*

1. **Polythetic divisive hierarchies for mixed symbolic data**
   Lynne Billard and Jaejik Kim
2. **Scientific collaboration: Comparing self evaluation and administrative data source**
   Luka Kronegger and Anuška Ferligoj
3. **Panel data estimation in regressions for symbolic data: An application to the clustering of cultural entrepreneurial regimes**
   Andrej Srakar and Marilena Vecco
4. **Clustering European countries regarding causes of death**
   Aleša Lotrič Dolinar, Jože Sambt and Simona Korenjak-Černe

19.30 **Workshop Dinner *Gostilna Čad***

9.00–10.00 **Software** (EF) *Chair: Oldemar Rodriguez*

1. **Latest developments of the RSDA 2.0: An R package for Symbolic Data Analysis**
   Oldemar Rodríguez

2. **Individual credit rating with ensemble learning for symbolic data**
   Marcin Pełka

10.00–10.30 **Coffee Break**

10.30–12.30 **Applications, Networks** (EF) *Chair: Javier Arroyo*

1. **New insights in political and biological problems by symbolic discriminant analysis**
   Sónia Dias, Paula Brito, Paula Amaral and Adelaide Freitas

2. **Clustering of symbolic data with relational constraint: Demographic sex-age structures in US and Europe**
   Nataša Kejžar, Simona Korenjak-Černe and Vladimir Batagelj

3. **Functional regionalization and the analysis of symbolic data**
   Carlo Drago and Alessandra Reale

4. **Symbolic network analysis of bike sharing systems**
   Vladimir Batagelj

12.30–14.00 **Lunch *Restaurant EF***

14.00–15.00 **Round Table** (EF)

15.00 **Closing**

ABSTRACTS

General SDA

Introduction to the Symbolic Data Analysis framework and application to post clustering for comparing and improving clustering methods by the ”Symbolic Data Table” that they induce

Edwin Diday

CEREMADE Paris-Dauphine University, Paris, France
diday@ceremade.dauphine.fr

First we recall that Symbolic Data Analysis (SDA) is a way of thinking by classes in Data Science. We recall that classes of standard units are in SDA the new statistical units of higher level than the initial standard statistical units. In SDA, classes are considered as objects to be described in all their facets by ”symbolic data” taking care on their internal variability by staying close of the user language. Then we focus on different strategies of building a Symbolic data table from a standard data table by using: Clustering by partitioning (k-means, dynamic clustering), Clustering by fuzzy partitioning (EM, others), mixture decomposition of Copulas (”copula-EM” or ”copula-dynamic clustering”). Few words will be said also on how building classes at the second level (where the units are classes), by using Dirichlet models. Then, we give tools in order to measure the quality of the obtained symbolic data tables. By this way we can compare the different associated clustering methods and improve them. Finally, we show how to summarize the obtained symbolic data tables (by a symbolic data table of symbolic data tables) and then, we show how to visualize and compare them and their associated clustering methods, for example by an extension of PCA and Pyramids to symbolic data tables.

Intervals, histograms, quantiles, compositions. . . what on earth should I do?

Paula Brito

FEP & LIAAD INESC TEC, University of Porto, Porto, Portugal
mpbrito@fep.up.pt

In recent years, many measures and methods have been developed to analyze symbolic data, i.e., data representing variability inherent to the observations. In general, methods have been developed independently, assuming data is readily available. These data take the form of intervals or distributions over a given domain, which are to be analyzed as such. They may be obtained from the aggregation of microdata, which is certainly the most frequent case; other times ranges of values or lists of quantiles are directly available. The question then arises as to which approach would be better suited for the problem at hand. When individual records are merged to obtain the descriptions of the concepts, which are the statistical units to be analyzed, the researcher often asks him/herself the question of whether to aggregate microdata in the form of intervals or histograms, or just compute an a priori given list of sample quantiles - and in that case how many. What are the questions here? Are there some assumptions behind each particular choice? Can we propose some guidelines? In this talk we shall review such issues, as well as the different options that lie ahead for the multivariate data analysis of interval and distributional data.

Meta-analysis and SDA

Masahiro Mizuta

Information Initiative Center, Hokkaido University, Sapporo, Japan

mizuta@iic.hokudai.ac.jp

The applicable range of SDA is expanding because SDA allows flexibility of data structure. In my presentation, I will discuss meta-analysis as an important application of SDA. Meta-analysis is a well-known technique in the fields of medical and social sciences. For example, summary statistics of several RCTs (randomized clinical trials) are integrated and analyzed in pharmacoepidemiology area. Source data of summary statistic in each RCT are picked up from the published papers through systematic review. This situation is exactly the framework of symbolic data analysis. Each scientific study or RCT can be considered a concept. We can find outliers and heterogeneity of scientific studies using symbolic clustering and symbolic discriminant analysis in meta-analysis.

Likelihood Functions

Estimating a mixture of Dependent Dirichlet Distributions

Richard Emilion

University of Orléans, Orléans, France
richard.emilion@univ-orleans.fr

A symbolic data analysis (SDA) table T with n rows and p columns of probability vectors of size $m_1, \ldots, m_p \in \{2, 3, \ldots\}$, respectively, is a good example of dependent multivariate blocks of variables. Mixtures of Dirichlet distributions (when $p = 1$) and mixtures of Dependent Dirichlet distributions (when $p \geq 2$) are proposed as parametric distributions that could generate T. Mixtures of Dirichlet Kernels are proposed in the nonparametric case. We investigate the estimation of such mixtures.

Probabilistic and statistical setting

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a fixed probability space. Let $M_1(\mathbb{V}_j)$ denote the space of all probability measures on a complete and separable metric space $\mathbb{V}_j$, $j = 1, \ldots, p$. Consider p random distributions $RD_j : \Omega \longrightarrow M_1(\mathbb{V}_j)$. For each j let $(V_{j,1}, \ldots, V_{j,l}, \ldots V_{j,m_j})$ be a fixed measurable partition of $\mathbb{V}_j$. As RD_j is not easy to handle, we rather consider the random probability vector

$$X_j = (RD_j(V_{j,1}), \ldots, RD_j(V_{j,l}), \ldots RD_j(V_{j,m_j})) \tag{1}$$

so that $X = (X_1, \ldots, X_j, \ldots, X_p)$ is a vector of p random probability vectors. The above data table T is then considered as an outcome of a n-sample of X.

Our setting covers several usual cases in SDA. If $RD_j(\omega)$ is a uniform distribution, RD_j is an interval-valued variable. Both cases of interval-valued variables and histogram-valued variables are covered by considering a partition $(V_{j,l})_l$ into adjacent non-overlapping intervals. If $\mathbb{V}_j = \{V_{j,1}, \ldots, V_{j,l}, \ldots V_{j,m_j}\}$ is a finite set of m_j elements, we cover the case of distributions of categorical variables.

Estimation procedures

When $p = 1$, we propose, as a model of distributions that can generate T, a finite mixture

$$\sum_{h=1}^{K} \lambda_h Dd(\alpha_h), \quad \lambda_h \geq 0, \quad \sum_{h=1}^{K} \lambda_h = 1 \tag{2}$$

of K Dirichlet distributions $Dd(\alpha_h)$ with parameter $\alpha_h = (\alpha_{h,1}, \ldots, \alpha_{h,l}, \ldots, \alpha_{h,m_1})$, respectively. As $Dd(\alpha_h)$ belongs to the exponential family, mixture (2) can be estimated using EM algorithm. The E (Expectation) step easily yields the estimation of the weights λ_h, the M (Maximization) part requires a numerical method to solve the system of equations

$$\sum_{i=1}^{n} t_{h,i}(\Phi(\sum_{l=1}^{m_1} \alpha_{h,l}) - \Phi(\alpha_{h,l}) - \log x_{i,1,l}) \;\; l = 1, \ldots, m_1. \tag{3}$$

When $p \geq 2$, from a centered Gaussian vector of size $m = m_1 + \ldots + m_p$ with p dependent blocks of independent real Gaussians, we get p dependent blocks of independent Gamma variables. Then dividing each block by the sum of its elements, we get p dependent Dd's. A mixture of Dependent Dd's is estimated using a variant of EM algorithm. Kernels and mixture of Kernels based on a Dd (resp. on a Dependent Dd) are proposed as nonparametric estimators.

A general framework for constructing symbolic likelihood functions

Scott A Sisson, Boris Beranger and Huan (Jaslene) Lin

University of New South Wales, Sydney, Australia

Scott.Sisson@unsw.edu.au, B.Beranger@unsw.edu.au, jaslenelin@hotmail.com

Symbolic data analysis is a convenient way to concisely summarise and then analyse large and complex datasets. However, very little work has focused on likelihood based inference, a mainstay statistical framework, and these focus on fitting models at the symbol level. Here we propose a general framework for constructing symbolic likelihood functions for fitting models to the underlying classical data, but where these data are only observed through the given symbols. This framework allows for the fitting of arbitrarily complex classical models to general forms of symbolic data, and moves away from common restrictive assumptions such as uniformity within intervals. We show that several existing classical data results involving random intervals and histograms can be derived as special cases, and demonstrate our approach on several simulated and real analyses.

Estimating species abundance using symbolic data meta-analysis

Huan (Jaslene) Lin[1], *Julian Caley*[2] *and Scott Sisson*[3]

[1] University of New South Wales, Sydeny, Australia

[2]Australia

[3] University of New South Wales, Sydney, Australia

jaslenelin@hotmail.com, julian.caley@gmail.com, scott.sisson@unsw.edu.au

Global species richness is a key biodiversity metric for which there has been a growing concern over its decline as a result of overexploitation and habitat destruction by humans. Despite an increasing awareness and efforts to estimate global species richness, the estimates in the literature both come with great uncertainty and are often logically inconsistent (in that estimates "lower down the tree" are often larger than those above it). Further, these estimates are typically represented in inconsistent forms, either as intervals (a, b) or point estimates with no uncertainty, which makes it difficult to combine information across different estimates. In this paper, we develop a Bayesian hierarchical approach to estimate the abundance of the number of species that naturally combines 45 estimates from previous studies in the literature. The data mix of intervals and point estimates are naturally reconciled using techniques from symbolic data analysis. This approach allows us to recover interval estimates at each species level, even when data is partially or wholly unobserved, while respecting logical constraints, and to determine the effects of estimation on the whole hierarchy of obtaining future estimates on particular species levels.

A composite likelihood based approach for max-stable processes using histogram-valued variables

Thomas Whitaker, Boris Beranger and Scott Sisson

University of New South Wales, Sydeny, Australia
t.whitaker@unsw.edu.au, b.beranger@unsw.edu.au,
scott.sisson@unsw.edu.au

The intractibility and computational intensity associated with large dimensional datasets means that classical methods of analysis are often unsuitable in the extraction of meaningful statistics. In recent years, Symbolic Data Analysis (SDA) has been introduced as a means of addressing such issues, whereby large datasets are summarised through the use of symbolic-valued variables such as intervals, histograms and distributions, which are then analysed in place of the classical-valued data. The analysis of spatial extremes using max-stable processes provides one such application of SDA. While classical composite likelihood methods appear to bypass the intractibility issues associated with the multivariate densities of max-stable processes with a reasonably large number of spatial locations, the computational cost of these methods is still too high when the spatial or temporal dimension gets too large. Here we introduce a symbolic composite likelihood approach to bypass these issues, whereby large spatial extremes datasets are aggregated into multidimensional histogram-valued symbols with random counts, leading to a reduction of the complexity of the data. The performance of our procedure in terms of inferential and computational efficiency is examined in an extensive simulation study and the impact of coarsening the data and the design of the symbols (histograms) is discussed. One result that remains necessary throughout is the recovery of the classical analysis that accompanies the convergence of the symbolic-valued data towards the classical case. Finally, the utility of the method is illustrated through the analysis of fortnightly maximum temperatures at 105 locations across Australia using historical data and simulated data from two climate models.

PCA, FA

Symbolic covariance for interval data: Properties and interpretation

M. Rosário Oliveira, Margarida Vilela and António Pacheco

CEMAT and Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
rosario.oliveira@tecnico.ulisboa.pt,
margarida.azeitona@tecnico.ulisboa.pt,
antonio.pacheco@math.tecnico.ulisboa.pt

Symbolic Data Analysis (SDA) consists on a set of methodologies designed to deal with complex structures of data existing on their own right or resulting from the aggregation of a base dataset according to the researchers' interest. Among the most basic needs to analyse these datasets are symbolic descriptive statistics like measures of location, dispersion, and association. The knowledge on their population counterparts and associated theoretical properties are important for the analyst do take full profit of these sample quantities.
In this work, we consider symbolic interval-valued variables and based on the available proposals of sample symbolic mean, variances and covariances we formulate its population counterparts. The theoretical properties of the population symbolic means, covariances, and correlations are demonstrated. In particular, we prove that the several proposals of symbolic correlations are quantities ranging between -1 and 1, similarly to what happens in the conventional framework.
We discuss micro-data models associated with each symbolic covariance definition under study. These models bring practical understanding and guidance for the practitioner to choose, among the available possibilities, the one that best fits a given dataset and consequently, to decide by a certain symbolic covariance definition.
Several multivariate SDA models are based on symbolic covariance matrices. For example, the majority of Symbolic Principal Component Analysis models for interval-valued variables mainly differ on the covariance structure considered. Given its importance, a unifying definition of population covariance and correlation matrices is introduced and their theoretical properties are derived. These results turn to be helpful to discuss the characteristics of some of the SPCA for interval-valued variables and in which way they relate to each other.
The analysis of a real dataset exemplifies the methodology of modelling micro-data by choosing a symbolic covariance definition to describe the variables associations and interpreting the estimated associations between the symbolic interval-valued variables under study.

Maximum likelihood factor analysis of interval data

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

[1]Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Viana do Castelo, & LIAAD INESC TEC, Universidade do Porto., Viana do Castelo, Portugal

[2]Faculdade de Economia & LIAAD INESC TEC, Universidade do Porto, Porto, Portugal

[3]Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Porto, Portugal

paulacheira@estg.ipvc.pt, mpbrito@fep.up.pt, psilva@porto.ucp.pt

Factor analysis models the dependence structure among a large set of variables by a small number of unobservable factors that explain the observed correlations. We propose a factor analysis model for interval-valued data, assuming a multivariate Normal distribution for the midpoints and log-ranges of the p interval variables. The method describes the correlation structure among the midpoints and log-ranges of the measured interval-valued variables in terms of a few underlying uncorrelated real-valued variables. Factors are extracted by the maximum likelihood method on the 2p $\times$ 2p midpoints and log-ranges correlation matrix. The identified factors capture possible relations among location, conveyed by the interval midpoints, among internal variation, conveyed by the interval log-ranges, and between location and internal variation. The method is evaluated on synthetic data with predefined correlation structures, and is applied to meteorological data.

Optimized symbolic principal component for interval-valued variables

Jorge Arce[1] *and Oldemar Rodríguez*[2]

[1]National Bank of Costa Rica, San José, Costa Rica

[2]University of Costa Rica, San Pedro, Costa Rica

jarceg@bncr.fi.c, oldemar.rodriguez@ucr.ac.cr

In the last two decades, principal component analysis was adapted for symbolic data, first in the context of interval-valued data. A number of approaches have been proposed. In Diday E. (1997) and Billard L. (2011), the authors proposed the centers method and the vertices methods to extend the well known principal components analysis method to a particular kind of symbolic objects characterized by multi–valued variables of interval type. Two methods were proposed, a vertices method which uses all the vertices of the observation's hypercube, and a centers method which uses the centroid values.

This paper aims to improve the centers method applying an optimization algorithms in which instead of projecting the centroid value we look for the best point to project in supplementary the vertices. The best point in the sense that it minimizes the distance of the supplementary individuals to that point or the point that generates a principal components analysis with the best inertia in the first components and then from this projecting the vertices as supplementary elements. We obtain interval–valued symbolic principal components which recapture better the internal variation of the observations or maximizes the correlation measures between these principal components and the random variables and/or the observations themselves.

Besides, the reader may use all the methods presented herein and verify the results using the RSDA package written in R language, that can be downloaded and installed directly from CRAN, see Rodríguez, O. (2017).

Visualization

Matrix decompositions based on induced norms with application to interval data

Vartan Choulakian

Université de Moncton, Moncton, Canada
vartan.choulakian@umoncton.ca

Matrix factorization, named also decomposition, in data analysis is at the core of factor analysis; and one of its principal aims is to visualize geometrically the statistical association existing among the rows or the columns of the matrix. So the way that we factorize a matrix is of fundamental interest and concern in statistics. The aim of this talk is : First, to present general theory of matrix factorization based on induced norms; then apply it to interval data.

A visualization synthesis through symbolic data analysis of open ended questions

Mireille Gettler Summa[1], *Myriam Touati*[1] *and Sadika Rjiba*[2]

[1]CEREMADE UMR CNRS, Paris Dauphine University, Paris, France
[2]Economics and Management Department, Sousse University, Sousse, Tunisia
summa@ceremade.dauphine.fr

We explore the potential and the efficiency of Symbolic Data Analysis on textual data coming from open-ended questions of surveys. Traditional textual analysis methods can visualize and resume the meaning of a text. Moreover clustering algorithms are developed to visualize relationships between words in order to structure and thus simplify texts; nevertheless they do not specify enough the relationships among the words constituting the concerned text. To overcome this problem, we first choose specific keywords that represent the core of the discussed topic. Data are eventually embedded in a fuzzy framework: each word of the corpus is given a membership degree that expresses the importance level of a word according to each keyword. Membership degree can be computed through fuzzy clustering approaches. Obtaining a high membership degree confirms the high contextual relationship between a keyword and any word from the context. We then mathematically formalize a keyword in the framework of Symbolic Data Analysis: its description consists in a finite discrete distribution of words in a given context. We assume that the possible underlying probability distributions may be different for each keyword according to each context. Because of this assumption we use Symbolic Data Analysis theoretical results and corresponding algorithms in order to visualize and structure the keywords according to their contexts. The contexts are the bags of words resulting from open-ended questions. Symbolic Factorial Analysis allow for the visualization of relationships among statistical units according to the free texts. The relevance of vocabulary lemmatization is questioned all along. We apply the approach on a Tunisian survey that includes open-ended questions.

Time Series

Data quality and hydrologic interval time series analysis

Carlo Drago, Silvia Di Francesco and Veronica Ciccone

University of Rome "Niccolò Cusano", Rome, Italy
carlo.drago@unicusano.it

Statistical analysis is very relevant in hydrological science. The use of the interval time series allows to focus on an higher range of problems but at the same time to take in to account the information provided by large datasets. Unfortunately classical time series and interval time series in hydrology are often characterized by errors and missing observations. A high quality of the data considered is an important condition to be able to conduct hydrologic statistical analyses. The problem is highly relevant for symbolic data and in particular for interval time series. In this work we present some statistical approaches on interval time series, considering both real and simulated data in order to overcome this problem.

Forecasting wind speed distributions with kernel regression

Albert Meco and Javier Arroyo

Facultad de Informática. Universidad Complutense de Madrid, Madrid, Spain
albert.meco@materiaworks.com, javier.arroyo@fdi.ucm.es

Wind speed is a continuous measure that is usually recorded at short time intervals as average values. However, wind speed can suffer great fluctuations in very short time and these are crucial for wind power generation. Many authors have already used distributions to characterize the wind speed fluctuations, however the use of distributions to characterize wind speed from the perspective of symbolic data analysis may render new insights into the phenomena.
In this work, we will forecast time series of wind speed distributions using kernel regression. Kernel regression is an instance-based learning approach that seems appropriate for wind speed forecasting. First, it is suitable to smoothly approximate non-linear functions and time series with different regimes, e.g. high and low variability. Second, it is also suitable for dealing with pattern-like behavior over time. These features are expected to be present in wind speed time series, hence interesting results can be anticipated.
We will consider time series of wind speed distributions measured at different frequencies in order to compare its predictability at different time scales. Furthermore, we will compare the results with those obtained by a naïve approach and the k Nearest Neighbours approach, other instance-based approach simpler than kernel regression. We will discuss the results obtained from these methods from different perspectives, including its ability to forecast distribution features, such as central tendency and variability, and computational aspects.

Applications, Clustering

Polythetic divisive hierarchies for mixed symbolic data

Lynne Billard[1] *and Jaejik Kim*[2]

[1]University of Georgia, Athens GA, USA
[2]Sungkyunkwan University, Seoul, Korea
lynne@stat.uga.edu, jaejik@skku.edu

We describe a polythetic divisive clustering algorithm for mixed data. By mixed data, we mean that some variables are modal/non-modal list data, some are interval data, and some are histogram data. The algorithm utilizes dissimilarity matrices for mixed symbolic data. An example is given for a Census data set.

Scientific collaboration: Comparing self evaluation and administrative data source

Luka Kronegger and Anuška Ferligoj

University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
luka.kronegger@fdv.uni-lj.si, anuska.ferligoj@fdv.uni-lj.si

Many studies have shown that measuring scientific collaboration by co-authorship covers only partially the scientific collaboration. Therefore, different other approaches to measure scientific collaboration were proposed, e.g. by surveys (Iglič et al. 2017) or mix method approach (Cimmino et al. 2017). It has been shown that the differences in the collaboration between disciplines are smaller when measured by a survey compared to co-authorship (Iglič et al. 2017). In this presentation we compare the structure of the selected scientific disciplines obtained by two types of measurement of the scientific collaboration.
In the analysis we explore 10 scientific disciplines in Slovenia (Physics, Mathematics, Biotechnology, Sociology, Economics, Materials, Neurobiology, Plant production and Histography), focusing on two independent measurements of scientific collaboration among Slovenian researchers. The first measurement is based on self evaluation of researchers gathered by a survey, the second one is made on data from national bibliographic database. Self evaluation of scientific collaboration and co-authorship are further divided to the collaboration inside a research group, research organization, inside national boarders and international collaboration. Self evaluation and co-authorship measure scientific collaboration on the individual level. Obtained variables are transformed into symbolic data – distributions of specific collaborations measured on the level of scientific discipline.
The aim of the analysis is to cluster 10 selected scientific disciplines according to two types of measurement of scientific collaboration and to compare the differences in distributions on the level of disciplines as well as on the whole dataset.

Panel data estimation in regressions for symbolic data: An application to the clustering of cultural entrepreneurial regimes

Andrej Srakar[1] *and Marilena Vecco*[2]

[1]Institute for Economic Research, Ljubljana and Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia

[2]Erasmus University Rotterdam, Rotterdam, The Netherlands

andrej_srakar@t-2.net, mari.vecco@gmail.com

"Entrepreneurial regimes" is a topic, receiving quite a lot of research attention in the recent years. As stated by some authors, for a more complete understanding of how entrepreneurship contributes to economic and societal development, it is important to recognize the contextually embedded quality of entrepreneurial actions and behaviors in national, regional, and city-level contexts. But, the topic has been seldom applied to the field of cultural entrepreneurship. Moreover, the existing studies on entrepreneurial regimes mainly use common methods from multivariate analysis (e.g. factor analysis, PCA, cluster analysis) and some type of institutional related analysis. In our analysis we study the cultural entrepreneurial regimes applying a symbolic data analysis approach and using Amadeus data for the period 2006-2015 and for 28 EU countries. Amadeus is a database, including comprehensive information on around 21 million companies across Europe and has to our knowledge not received many applications using symbolic data analysis to date. On the basis of these data (and some other secondary sources) we extract a set of histogram variables related to characteristics of the firms in culture (socio-demographical and socio-economic characteristics and financial variables and ratios) and general features of their macroeconomic environment (e.g. GDP per capita, employment, inflation, public expenditure, public debt, gross savings and investment). These variables are employed in the symbolic clustering analysis, following the most recent trends, to derive a set of temporally robust clusters, labelled as cultural entrepreneurial regimes. Next, we try to derive mathematical formulas for panel data estimators for regression analysis with symbolic data (to our knowledge not done so far in the existing literature) and explore their statistical (asymptotic) behaviour. Finally, the behaviour of clusters (cultural entrepreneurial regimes) is analyzed in regressions for symbolic data, including panel data estimation. The main research questions of the article are: 1) To what extent do the clusters (cultural entrepreneurial regimes) follow the commonly found classifications of entrepreneurial regimes (following the literature mentioned before)? 2) Are there any significant changes in the positions of individual countries observed in the studied time period? 3) What are the explanations for these changes? 4) Does the inclusion of formal panel data modelling improve results from more common regression (e.g. POLS) specifications for symbolic data in the studied empirical example? The analysis is to our best knowledge the first symbolic data analysis in cultural entrepreneurship and economics, focusing specifically on the symbolic nature of the variables. Moreover, it derives formulas for panel data estimators in the case of regressions for symbolic data, contributing to the development of symbolic data analysis research field.

Clustering European countries regarding causes of death

Aleša Lotrič Dolinar, Jože Sambt and Simona Korenjak-Černe

Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia
alesa.lotric.dolinar@ef.uni-lj.si, joze.sambt@ef.uni-lj.si, simona.cerne@ef.uni-lj.si

People in countries with different sex-age specific mortality suffer from different health problems and consequently the countries face different health costs. For implementing proper health policy, it is important to know which groups of countries are similar and what are the differences between the groups. In this paper we study the mortality in European countries using classical and SDA clustering methods, based on different types of input data: death level only, relative structure of deaths by causes of death, and combination of the two. As causes of death are strongly related to sex and age, the input data are provided separately for each sex-age group. The clustering results are compared based on how they capture both the level of mortality and the relative distribution of deaths by causes of death. We compare the results also with the clusters based on life expectancy at birth that is considered as the main demographic indicator of country development.

Software

Latest developments of the RSDA 2.0: An R package for Symbolic Data Analysis

Oldemar Rodríguez

University of Costa Rica, San Pedro, Costa Rica
oldemar.rodriguez@ucr.ac.cr

This package aims to execute some models on Symbolic Data Analysis. Symbolic Data Analysis was propose by the professor E. Diday in 1987 in his paper "Introduction à l' approche symbolique en Analyse des Données". Premiére Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987. A very good reference to symbolic data analysis can be found in "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis" of L. Billard and E. Diday that is the journal American Statistical Association Journal of the American Statistical Association June 2003, Vol. 98.
The main purpose of Symbolic Data Analysis is to substitute a set of rows (cases) in a data table for an concept (second order statistical unit). For example, all of the transactions performed by one person (or any object) for a single "transaction" that summarizes all the original ones (Symbolic-Object) so that millions of transactions could be summarized in only one that keeps the customary behavior of the person. This is achieved thanks to the fact that the new transaction will have in its fields, not only numbers (like current transactions), but can also have objects such as intervals, histograms, or rules. This representation of an object as a conjunction of properties fits within a data analytic framework concerning symbolic data and symbolic objects, which has proven useful in dealing with big databases.
In RSDA version 2.0, methods like centers interval principal components analysis, histogram principal components analysis, multi-valued correspondence analysis, interval multi-demensional scaling (INTERSCAL), symbolic hierarchical clustering, CM, CRM, Lasso, Ridge and Elastic Net Linear regression model to interval variables have been implemented. This new version also includes new features to manipulate symbolic data through a new data structure that implements Symbolic Data Frames and methods for converting SODAS and XML SODAS files to RSDA files. This version also includes Optimized Center Method and Principal Surfaces to Principal Component Analysis and new plot graphics like radar charts to interval variables.

Individual credit rating with ensemble learning for symbolic data

Marcin Pełka

Wrocław University of Economics, Wrocław, Poland
marcin.pelka@ue.wroc.pl

A credit rating means an evaluation of the credit risk for an individual, company or debt security (for example a bond). Usually credit rating is built upon the basis of the credit history, present financial condition, future income, etc. Usually methods like logistic regression, multivariate adaptive regression splines (MARS), decision trees are used for credit scoring. However also ensemble methods like bagging, boosting or random forests can be useful in credit scoring. In classical data analysis objects (patterns) are usually described by single valued variables. This allows to represent them as a vector of qualitative or quantitative measurements, where each column represents a variable. However this kind of data representation is too restrictive to represent more complex data. To take into account the uncertainty and/or variability into account the uncertainty and/or variability to the data, variables must assume sets of categories or intervals, even with frequencies or weights. Such kind of data have been mainly studied in Symbolic Data Analysis (SDA). The paper presents an application of random forests and bagging (built upon decision trees only, or decision trees, kernel discriminant analysis and k-NN methods) methods for credit rating for symbolic data. SymbolicDA package of R software and R software were used for all calculations. The results obtained from ensemble models were compared with a single model of the same type (decision tree, k-NN, kernel discriminant analysis). All simulations show that ensemble learning for symbolic data allows to get better prediction (when considering adjusted Rand index) than a single model and all proposed solutions can be used for an individual credit scoring task.

Applications, Networks

New insights in political and biological problems by symbolic discriminant analysis

Sónia Dias[1], *Paula Brito*[2], *Paula Amaral*[3] *and Adelaide Freitas*[4]

[1]School of Technology and Management, Polytechnic Institute of Viana do Castelo and LIAAD/INESC-TEC, University of Porto, Viana do Castelo, Portugal

[2]Faculty of Economics and LIAAD INESC-TEC, University of Porto, Porto, Portugal

[3]CMA and Faculty of Science and Engineering, University Nova de Lisboa, Lisboa, Portugal

[4]CIDMA and Department of Mathematics of the University of Aveiro, Aveiro, Portugal

sdias@estg.ipvc.pt, mpbrito@fep.up.pt, paca@fct.unl.pt, adelaide@ua.pt

In this presentation we analyze real applications of discriminant analysis based on a linear combination of distributions or intervals. The proposed approach uses a linear discriminant function defined according to the Distribution and Symmetric Distribution linear regression model, where distributions or intervals are represented by quantile functions, under some assumptions. This discriminant function allows defining a score for each individual, in the form of a quantile function. Classification in two a priori groups is then based on the Mallows distance between the score of the individual and the score obtained for the barycentric histogram of each a priori class.

There is a diversity of application areas for the proposed linear discriminant method. In this work we investigate problems arising in two different areas.

The first one concerns social/political data. After an electoral act between two candidates or political parties, the maps of the countries may be colored accordingly, and it is then possible to know who the winner is in each region or state. On the other hand, we may aggregate census data on social and economic characteristics of the population by state or region. From the resulting distributional variables, and applying the proposed linear discriminant method, it is then possible to determine the variables that allow for the characterization of the regions that voted for each of the candidates or political parties. This approach is applied to presidential American elections.

The second problem concerns genomic data of twenty three bacterial species, of which eleven are Gram Positive and twelve are Gram Negative. The complete genome of a species consists of thousands of genes, each gene is a sequence of codons composed by three positions (nucleotides), where each position is occupied by one of the four nucleotides (A, T, G, C). It has been shown that the frequencies of each nucleotide or couple of nucleotides in each codon position, or in the three codon positions, allow classifying the bacteria in different types. In the proposed approach, the frequencies of the four nucleotides are aggregated by species. The goal is to investigate whether the characteristics at the gene level, i.e., considering the distribution of frequencies of each nucleotide by gene, for each genome/species, allows separating the bacteria as Gram Positive or Gram Negative.

Clustering of symbolic data with relational constraint: Demographic sex-age structures in US and Europe

Nataša Kejžar[1], *Simona Korenjak-Černe*[2] *and Vladimir Batagelj*[3]

[1]University of Ljubljana, Faculty of Medicine, Ljubljana, Slovenia
[2]University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia
[3]IMFM Ljubljana and UP IAM Koper, Ljubljana, Slovenia
natasa.kejzar@mf.uni-lj.si, simona.cerne@ef.uni-lj.si,
vladimir.batagelj@fmf.uni-lj.si

In this work we combine clustering of symbolic data units with clustering with relational constraints. We propose an adapted hierarchical clustering method (Batagelj et al, arXiv 2015, Ferligoj and Batagelj, 1982, 1983). For an illustration it is applied on maps and population pyramids of (a) United States counties from 2006 and (b) of European Nomenclature of territorial units for statistics (NUTS) from 2013.
Population pyramids represent symbolic objects (SOs) with two modal-valued variables. From a map we get geographical adjacency of units (SOs) which imposes their relational constraints. The feasible clusterings of SOs are those that induce a connected subgraph – they form a geographically contiguous group of units.

Functional regionalization and the analysis of symbolic data

Carlo Drago[1] *and Alessandra Reale*[2]

[1]University of Rome "Niccolò Cusano", Rome, Italy
[2]ISTAT - Italian National Institute of Statistics , Rome, Italy
c.drago@mclink.it, reale@istat.it

The analysis of geographical and economical spaces it could be important to divide a territory in different partitions or regions. Community detection algorithm are very useful to identify the different zones belong a spatial network. In this sense it could be important by considering the structure of the network and then collapsing the nodes belong a single community. These zones are characterized by different structural and social descriptors. Collapsing the nodes leads naturally to consider a specific interval as the representation of the spatial zone identified. In this context we can to represent the different variables obtained as descriptors of the different regions or zones as symbolic data which allows to keep into account the variability of the original nodes on the spatial network. At the same time we are interested on analyzing the different zones and so we conduct an analysis of the symbolic data representing the different zones. In order to consider the linkages between the different zone we are able to detect the inter-communities structures. The inter-community structure is important to understand the relationships between the different zones or regions of the network. This structure is related to the links connecting different communities. The result is also relevant because it is possible in this sense considering at the same time the flows between the different nodes and so the flows between the different regions.

Symbolic network analysis of bike sharing systems

Vladimir Batagelj

IMFM Ljubljana and UP IAM Koper, Ljubljana, Slovenia
vladimir.batagelj@fmf.uni-lj.si

Many cities around the world provide a bike sharing service. Some of them (San Francisco Bay Area, New York, Chicago, Boston, Philadelphia, Washington D.C., Melbourne, Paris, London, etc.) offered as open data the data sets about trip histories. They usually consist of the trip start and end (day, time, station) and member's or bike's id. Sometimes they are augmented by additional data about stations and changes of their status. For some data sets also the weather data were collected. On the basis of these data, interesting insights into the dynamics of bike sharing systems can be obtained.
We propose different symbolic networks that summarize selected aspects of a given bike sharing system. We present some analyses of selected bike sharing systems based on them.

INDEX OF AUTHORS

Index of Authors

SUPPORTED BY

Institute of Mathematics, Physics and Mechanics
http://www.imfm.si/

Statistical Society of Slovenia
http://www.stat-d.si/

SURS – Statistical Office of Republic of Slovenia
http://www.stat.si/

University of Ljubljana, Faculty of Economics
http://www.ef.uni-lj.si/