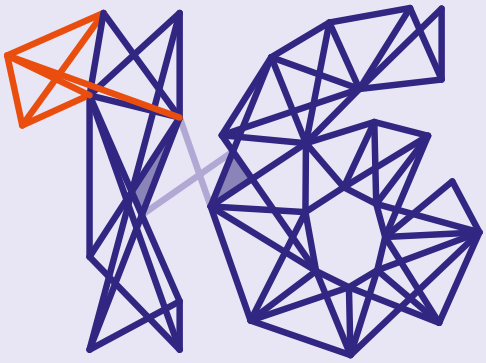


Abstract Book

16th Conference of the International Federation of Classification Societies

26-29 August 2019
Thessaloniki Concert Hall
Thessaloniki, Greece
#IFCS2019

ORGANISED BY

ifcs

LOCAL ORGANISER

IN COOPERATION WITH

THE GREEK SOCIETY OF DATA ANALYSIS

DIRECTOR OF THE CONFERENCE
Theodore Chadjipadelis
T. +30 2310 997912, **E.** chadji@polsci.auth.gr

16th Conference of the International Federation of Classification Societies

26-29 August 2019
Thessaloniki Concert Hall
Thessaloniki, Greece
#IFCS2019

ORGANISED BY

ifcs

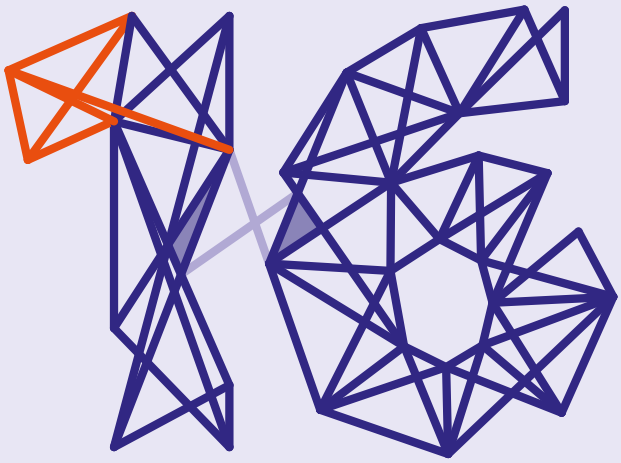
LOCAL ORGANISER

IN COOPERATION WITH

THE GREEK SOCIETY OF DATA ANALYSIS

ISBN: XXX-XXX-XX-XXX-X-X
© Copyright 2019, The Organizing Committee of the 16th Conference of the International Federation of Classification Societies
Publications Management: ARTION Conferences & Events

TABLE OF CONTENTS



CHAIRMAN'S WELCOME LETTER

Dear colleagues,

We are pleased to inform you that the 2019 conference of the International Federation of Classification Societies (IFCS) will take place on 26th - 29th of August 2019 in Thessaloniki (Greece).

On behalf of the Thessaloniki local organisers (Aristotle University of Thessaloniki in cooperation with Greek Society of Data Analysis – GSDA) and on behalf of the Federation we would like to invite you to join the IFCS-2019 Conference. The International Federation of Classification Societies (IFCS) founded in 1985 and is composed of many statistical societies all over the world. IFCS is an interdisciplinary and international organisation whose main purposes are to promote the scientific study of classification and clustering (including systematic methods of creating classifications from data), and to disseminate scientific and educational information related to its fields of interests.

The conference will bring some of leading researchers and practitioners in the related areas and will provide an opportunity for exchanging ideas, between researchers and practitioners, and establishing networking and collaborations. The conference will include several invited papers on important and timely topics from well-known leaders in the field, and parallel tracks of oral presentation sessions of the accepted papers. We will be glad to welcome you all to Thessaloniki in August 2019!

Best Regards,

Theodore Chadjipadelis

Director of the IFCS-2019 Conference

IFCS 2019

The 2019 conference of the International Federation of Classification Societies (IFCS) will be held in Thessaloniki, Greece from August 26 to 29, 2019. The conference theme will be 'Data Analysis and Rationality in a Complex World'. The conference opening will take place on August 26 late afternoon and pre-conference workshops will be held. The conference itself will start on August 27 in the morning, and will close on August 29 with a full day conference program and a conference dinner. The conference will include a President's invited session and a Presidential address, invited presentations, invited and contributed sessions, and poster sessions. The following awards and medals ceremony will be scheduled at the conference.

Helga and Wolfgang Gaul Stiftung Award (age < 30)

Chikio Hayashi Award (age 30 – 35)

IFCS Research Medal for outstanding research achievements (age > 35)

Student/Postdoctoral Fellow Paper Competition and Travel Award (age 22 – 29)

Cluster Benchmarking Competition Award

COMMITTEES

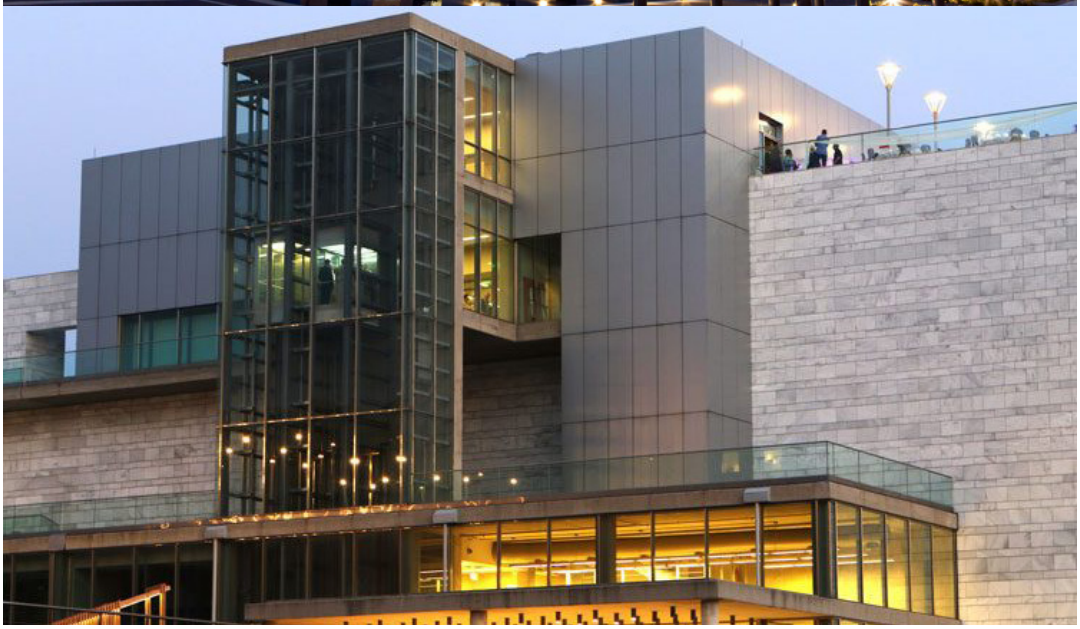
Scientific Program Committee

Theodoros Chadjipadelis	Chair
Berthold Lausen	Vice-chair, IFCS President
Tae Rim Lee	Vice-chair
Angela Montanari	IFCS President Elect
Akinori Okada	IFCS Past President
Christian Hennig	IFCS Scientific Secretary
Sugnet Lubbe	IFCS Treasurer
Katrijn van Deun	Publication Officer
Eva Boj	SEIO-AMyC
Paula Brito	CLAD
Sanjeena Dang	CS
Giannoula Florou	GSDA
Krzysztof Jajuga	SKAD
Hans Kestler	GfKl
Simona Korenjak Černe	SSS
Peter Kovacs	HSA-CMSG
Koji Kurihara	JCS
Angelos Markos	GSDA
Paul McNicholas	CS
Fionn Murtagh	BCS
Mauricio Vichi	CLADAG
Roberto Rocci	CLADAG
Mark de Rooij	VOC
Niel le Roux	SASA-MDAG

Local Organizing Committee

Theodore Chadjipadelis	LOC Chair, Aristotle University Thessaloniki
Giannoula Florou	President GSDA
Sofia Anastasiadou	University of Western Macedonia
Dimitris Karapistolis	Alexandrian Technological Institute
Nikos Koutsoupias	University of Macedonia
Angelos Markos	Democritus University of Thrace
George Menexes	Aristotle University Thessaloniki
Odysseas Moschidis	University of Macedonia
Thanos Thanopoulos	Hellenic Statistical Authority
Christos Tzimos	Hellenic Statistical Authority
Despoina Amarantidou	ARTION conferences & events
Vicky Bouranta	Secretary
Marina Sotiroglou	Secretary

Thessaloniki Concert Hall



The **Thessaloniki Concert Hall** (Building M2) is located along the coast of the city, close to the city center and the airport. Designed by the renowned architect Arata Isozaki, M2 building adorns the city with a unique construction that epitomizes the virtues of modern architecture. Geometrical lines, extended glass surfaces and elements of steel compose an image of imposing simplicity. Filled with natural light and enjoying a superb view to the sea, M2's foyer is a vast space of sophisticated aesthetics that expands in three levels and adjoins all of the Hall's important spaces. Equipped with cutting-edge technology and having an exceptional infrastructure, it provides a contemporary cultural and conference center of international standards with the capacity to host various events.

Get directions

CONFERENCE TOPICS

Contributed papers from scholars and practitioners are invited on any of the topics below as well as on related issues:

Data Analysis and Classification in the Natural Sciences, Engineering, Medicine and Biology / Banking and Finance / Bioinformatics and Biostatistics / Biomedical Data Analysis and Imaging / Business and Industry / Classification, Discriminant Analysis, and Supervised Learning / Clustering and Unsupervised Learning / Complex Event Processing / Data Analysis in Archaeology / Databionics / Exploratory Data Mining / Formal Concept Analysis / Functional Data Analysis / Geospatial Planning / Graphics and Vision / High-Dimensional Data and Dimension Reduction / Image Analysis and Computer Vision / Information Retrieval and Library Science / Impact of Technical Revolution, Globalization on Statistical Process / Knowledge Discovery / Linguistics and Musicology / Machine Learning and Pattern Recognition / Marketing and Management / Medical and Health Sciences / Neural Networks and Deep Learning / Psychology and Educational Sciences / Social Network Analysis / Sparse Modeling / Statistical Data Analysis, Statistical Models and Model Selection / Swarm Systems / Symbolic Data Analysis / Text Mining, Web Mining, and Ontology Learning / Visualization. Special sessions on Official Statistics will be organized in collaboration with the Hellenic Statistical Authority.

The Cluster Benchmarking Task Force of the IFCS is calling for contributions to a cluster benchmarking competition as part of IFCS-2019. Benchmarking studies will be presented in special sessions.

SOCIAL PROGRAM

On Monday August 26th at 8:30pm, after the conference opening, the welcome reception will take place at the Prefecture of Thessaloniki (Leof. Vasilissis Olgas 198), a 15min walk from the conference venue.

The conference dinner will be held on Wednesday August 28th at 8:30p.m., in a beautiful restaurant by the sea. In a currently renewed environment just beside the sea you can experience a perfect gastronomical dinner.

A bus city tour will be provided starting at 7.10p.m. from 25 Martiou Str., the main street in front of the conference venue. The bus will end up at the restaurant.

INVITED SPEAKERS

David J. Hand

is Emeritus Professor of Mathematics and Senior Research Investigator at Imperial College, London, where he formerly held the Chair in Statistics. He is a Fellow of the British Academy, and an Honorary Fellow of the Institute of Actuaries, and has served (twice) as President of the Royal Statistical Society. He is a non-executive director of the UK Statistics Authority, a member of the European Statistical Advisory Committee, a member of the International Scientific Advisory Committee of the Canadian Statistical Sciences Institute, a member of the Advisory Board of the ONS's Data Science Campus, and a member of the Advisory Board of the Cambridge Institute for the Mathematics of Information. He previously Chaired the Research Board of Imperial College's Data Science Institute and was Chair of the UK's Administrative Data Research Network. He spent eight years as Chief Scientific Advisor to Winton Capital Management. He has published 300 scientific papers and 29 books. In 2002 he was awarded the Guy Medal of the Royal Statistical Society, and in 2012 he and his research group won the Credit Collections and Risk Award for Contributions to the Credit Industry. He was awarded the George Box Medal in 2016. In 2013 he was made OBE for services to research and innovation.

Vladimir Batagelj

is Professor Emeritus of the University of Ljubljiana, Slovenia. He is also a member of the Institute of Mathematics, Physics and Mechanics (IMFM), Ljubljana, and of AMI, UP, Koper. His coauthored book *Generalized Blockmodeling* was awarded the 2007 Harrison White Outstanding Book Award by the Mathematical Sociology Section of the American Sociological Association. From the International Network for Social Network Analysis he was awarded the Georg Simmel Award (2007) and the Richards Software Award for the program Pajek (2013).

Theodoros Evgeniou

is a Professor of Decision Sciences and Technology Management at INSEAD in Fontainebleau France and an Academic Director of INSEAD eLab, a research and analytics center at INSEAD that focuses on data analytics for business. He graduated first in the MIT class of 1995 dual degrees in Mathematics, won medals in international mathematical Olympiads, and European awards for business case studies. At INSEAD, Theodoros has been focusing on data analytics applied to a range of areas from customer insights and marketing to finance. He has been developing and teaching courses on Data Analytics, Statistics and Decision Making.

Michael Greenacre

is Professor of Statistics at the Universitat Pompeu Fabra, Barcelona, Spain. He has authored and co-edited 10 books and over 100 journal articles and book chapters, mostly on *correspondence analysis* but more recently on *compositional data analysis.* He has given short courses in fifteen countries to environmental scientists, sociologists, data scientists and marketing professionals, and has specialized in statistics in ecology and social science. He is presently participating in many different research projects on Arctic ecology.

David Hunter

is Professor at Penn State Department of Statistics. His research interests include statistical computing, models for social networks, and statistical clustering. He has published extensively on networks, optimization algorithms and mixture models.

Julie Josse

is Professor of Statistics at Ecole Polytechnique in France. She has specialized in missing data, visualization and the nonparametric analyses of complex data structures. She has published over 30 articles and written 2 books in applied statistics. Julie Josse has developed packages to transfer her works such as missMDA dedicated to missing values. She is deeply involved in the R community and is part of Rforwards to widen the participation of minorities in the communities.

Andy Mauromoustakos

is Professor at the Department of Crop, Soil, and Environmental Sciences and works as an Applied Statistician for the AGRI STAT LAB at the University of Arkansas Fayetteville campus. Andy teaches graduate courses and does research related to Experimental Designs and Data Analysis for the AG Experiment Station and the Division of AG.

Sofia Olhede

is Professor of Statistics at University College London, director of UCL's Centre for Data Science, an honorary professor of Computer Science and a senior research associate of Mathematics at University College London. Sofia has contributed to the study of stochastic processes; time series, random fields and networks. She is on the ICMS Programme Committee since September 2008, a member of the London Mathematical Society Research Meetings Committee, a member of the London Mathematical Society Research Policy Committee and an associate Editor for Transactions in Mathematics and its Applications. Sofia was also a member of the Royal Society and British Academy Data Governance Working Group, and the Royal Society working group on machine learning.

PRE-CONFERENCE WORKSHOPS

Workshop 1: Compositional Data Analysis in Practice

Date: Monday, August 26, 2019, 08.30am - 11:30am
Venue: University of Macedonia (Central PC Lab, 1st floor)
Fee: 50EUR (paid via the conference website or on-site)

Name of instructor:
Prof. Michael Greenacre, Universitat Pompeu Fabara, Barcelona.

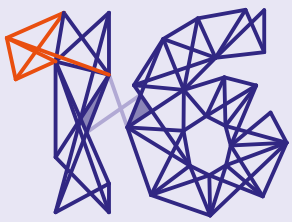
Short description:
Compositional data are multivariate data with the constant sum constraint, for example sets of nonnegative data that each sum to 1, often found in chemistry (samples in bio- and geochemistry), sociology and economics (time and monetary budgets) and marketing (market shares), for example. The key idea is the logratio transformation, which has certain implications for the analysis and then the interpretation of the results. This short course explains the main features of this novel area of statistics, with many illustrations to real data in a variety of contexts.

Introductory background:
Greenacre, M.J. (2018). Compositional Data Analysis in Practice. Chapman & Hall / CRC Press.
URL: https://github.com/michaelgreenacre/CODAinPractice

Tentative schedule:
1. Compositional data and the logratio transformation: practical implications and interpretation [50 min]
2. Total logratio variance and logratio distance; logratio analysis [50 min]
Break [20 min]
4. Modeling with logratios; variable selection; software [60 min].

Target audience:
Practitioners and researchers related to any domains where compositional data are found. Statisticians who want an introduction to this field of research and application.

Facilities required:
- Course participants should preferably bring their own laptops.
- Software: R, open source, with the R package **easyCODA** installed (which also requires the **ca**, **vegan**, **boot** and **ellipse** packages
- Course Material. All course materials, including the data and R scripts for the examples, will be made available for course participants.



Workshop 2: Symbolic Data Analysis: Parametric multivariate analysis of interval data

Date: Monday, August 26, 2019, 11.40am - 15.30pm
Venue: University of Macedonia (Central PC Lab, 1st floor)
Fee: 50EUR (paid via the conference website or on-site)

Names of instructors:
Paula Brito, University of Porto, Portugal
Pedro Duarte Silva, Católica Porto Business School, Portugal.

Short description:
Symbolic Data is concerned with analysing data with intrinsic variability, which is to be taken into account. In Data Mining, Multivariate Data Analysis and classical Statistics, the elements under analysis are generally individual entities for which a single value is recorded for each variable - e.g., individuals, described by age, salary, education level, etc. But when the elements of interest are classes or groups of some kind - the citizens living in given towns; car models, rather than specific vehicles - then there is variability inherent to the data. Symbolic data goes beyond the usual data representation model, considering variables whose observed values for each element are no longer necessarily single real values or categories, but may assume the form of sets, intervals, or, more generally, distributions. In this Tutorial we focus on the analysis of interval data, i.e., when the variables' values are intervals of IR, adopting a parametric approach. The proposed modelling allows for multivariate parametric analysis; in particular M(ANOVA), discriminant analysis, model-based clustering, robust estimation and outlier detection are addressed. The referred modelling and methods are implemented in the R package MAINT.Data, available on CRAN.

Introductory background:
Brito P. (2014). Symbolic data analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, *4*(4), 281-295.
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.

Tentative schedule:
1. Introduction to Symbolic Data Analysis: Motivation. Examples. Types of symbolic variables and their representations. Sources of symbolic data: aggregation of microdata. [60 min].
Break [15 min]
2. Parametric modelling of interval data: Gaussian and Skew-Normal models. (M)ANOVA, Discriminant Analysis, Robust estimation and outlier detection, Model-based clustering [90min].
Break [15 min]
3. Case-studies with R Package MAINT.Data [60 min].

Target audience:
The course is aimed at all potential data analysts who need or are interested in analyzing data with variability, e.g. data resulting from the aggregation of individual records into groups of interest, or data that represent abstract entities such as biological species or regions as a whole. This methodology is particularly interesting for Economics and Management studies, Marketing, Social Sciences, Geography, Official Data statistics, as well as for Biology or Geology Data Analysis.

It is assumed that the participants have a good background in classical Statistics and Multivariate Data Analysis.

Facilities required:
- Course participants' should bring own laptops, with R, RStudio, and the R package
- MAINT.Data installed.
- Course Material. All course materials, including the data and examples of software used for the case studies, will be made available for course participants.

INVITED SESSIONS

Theofanis Exadaktylos, Theodore Chadjipadelis (GSDA/ECPR): Analysis of European Parliament Elections
José Fernando Vera & Eva Boj del Val (SEIO-AMyC): New developments in clustering and scaling data
Christian Hennig (BCS): Philosophy relevant to classification and data science.
Christian Hennig (IFCS Cluster Benchmarking Task Force): Neutral Benchmarking Studies of Clustering
David Hunter (BCS): Statistical theory of cluster analysis
Salvatore Ingrassia (CLADAG): Advances in Mixture Modeling
Krzysztof Jajuga (SKAD): Data Analysis in finance
Aglaia Kalamatianou (GSDA): Data mining techniques and classification methods in Social Sciences
Nataša Kejžar, Simona Korenjak-Černe, Andrej Srakar (SSS): Advances in classification analysis for complex data – compositional and symbolic approaches
Koji Kurihara (JCS): Clustering for spatio-temporal data and its visualization
Paul McNicholas (CS): Clustering, Classification and Data Analysis via Mixture Models
Angela Montanari (CLADAG): Supervised classification with imprecise labels and complex data
Theophilos Papadimitriou, Periklis Gogas (GSDA): Emerging methodologies in economics and finance
Iannis Papadimitriou (GSDA): Developments of data analysis in Greece
Jozef Pociecha (SKAD): Classification methods in economics and business
Mark de Rooij (VOC): Crossroads of Statistical Learning and Psychometrics
Niel le Roux (SASA-MDAG): Classification, visualisation and dimension reduction
Cristina Tortora (GSDA): Clustering categorical and mixed-type data

President's invited session

Berthold Lausen, Theodore Chadjipadelis: Data Science Education

Presidential address

Berthold Lausen: Predictive ensemble methods for event time data

POST CONFERENCE PROCEEDINGS

Post conference proceedings will be published in Studies in Classification, Data Analysis, and Knowledge Organization. *(Edited by Theodore Chadjipadelis, Berthold Lausen, Angelos Markos, Tae Rim Lee, Angela Montanari, Rebecca Nugent).* Single papers can be made open access on payment of the open access to Springer.

THE CITY OF THESSALONIKI

The city was founded in 315 BC by Cassander, in honor of his wife Thessaloniki, sister of Alexander the Great. Since then, and due to its strategic position, Thessaloniki has been a commercial and cultural crossroad that brought together people and ideas from all over the world. The signs of this uninterrupted urban activity for more than 2,300 years are evident in each corner of the city. Nowadays, Thessaloniki is a big, modern city, with a population of around one million, and an important administrative and financial center of the Balkans. The warm and vibrant city life is largely influenced by the Aristotle University of Thessaloniki and the University of Macedonia which host tens of thousands of students. Thessaloniki is surrounded by places of great natural and historic beauty such as Olympus National Park, Vergina, where the Royal tomb of Philip II, father of Alexander the Great was found, the autonomous Mouth Athos, which is forbidden to women and children, and Halkidiki with its beautiful sandy beaches.

People & Life

Thessaloniki is a popular destination. You will certainly enjoy a pleasant and interesting stay in the city. People are friendly and happy to help with any questions. The atmosphere is unique during the day in the commercial and shopping centre, but especially during the evening, in the wide variety of bars, restaurants and theatres for entertainment. Thessaloniki is renowned for its unique location, along the Thermaikos Gulf, its sunsets, its long history, its monuments and museums as well as its distinguished cuisine.

IMPORTANT DATES

Monday, November 19, 2018	Start early bird registration
Sunday, May 05, 2019	Deadline for abstract submission
Sunday, May 19, 2019	Notification of acceptance for abstract submission
Sunday, May 26, 2019	Deadline for early bird registration
Sunday, June 23, 2019	Deadline for standard registration
Sunday, July 21, 2019	Deadline for late registration
Monday, August 26, 2019	Conference opening and pre-conference workshops
August 27-29, 2019	IFCS-2019 conference sessions

PUBLICATIONS RELATED TO THE CONFERENCE

We are planning to publish a post-conference proceedings. The tentative schedule is as follows:

December 31, 2019	Manuscript submission deadline
January 30, 2020	First notification
February 28, 2020	Deadline revised manuscript submission
June 15, 2020	Publication and shipment

MEETING SECRETARIAT

ARTION CONFERENCES & EVENTS
PROFESSIONAL CONGRESS ORGANISER FOR IFCS-2019 CONFERENCE

www.artion.com.gr

E. ifcs@artion.com.gr
T. +30 2310 257803 (direct line), +30 2310 272275
W. www.ifcs.gr

Conference Coordination
Despina Amarantidou

Co-ordination of the scientific program
Chara Ignatiadou, Kelly Angelaki

Management of delegates and residence
Markos Papadopoulos

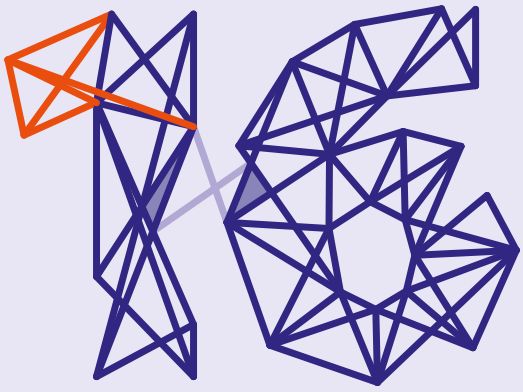
Marketing, Publications, Sponsors
Prodromos Nikolaidis, Lila Stathaki, Efi Mamoglou

IT
George Kanakaris

PROGRAM

Monday, 26th August 2019

	Venue: University of Macedonia (Central PC Lab, 1st floor)
08.00 – 08.30	Workshop registration
08.30 – 10.00	Workshop on **Compositional Data Analysis in Practice** Michael Greenacre (Part I)
10.00 – 10.20	**Coffee Break**
10.20 – 11.30	Workshop on **Compositional Data Analysis in Practice** Michael Greenacre (Part II)
11.40 – 13.00	Workshop on **Symbolic Data Analysis: Parametric multivariate analysis of interval data** Paula Brito and Pedro Duarte Silva (Part I)
13.00 – 14.00	**Lunch Break**
14.00 – 15.30	Workshop on **Symbolic Data Analysis: Parametric multivariate analysis of interval data** Paula Brito and Pedro Duarte Silva (Part II)
	Venue: Thessaloniki Concert Hall (Building M2)
16.30 – 18.00	Registration
18.00 – 18.30	Opening Ceremony — Plenary Hall
18.30 – 20.00	Panel discussion on **Data Science, Elections and Government** Athanasios Thanopoulos, *President of the Hellenic Statistical Authority* **Moderator:** Theodore Chadjipadelis
20.30	Welcome Reception



Tuesday, 27th August 2019

Venue: Thessaloniki Concert Hall (Building M2)

08.00 – 09.00 Registration

Parallel Sessions

09.00 – 10.40 **CONTR1: Supervised learning and applications I**
Chair: Pedro Duarte Silva — Room I

Multiclass posterior probability support vector machines for big data
Pedro Duarte Silva

Improving credit client classification by deep neural networks?
Klaus Bruno Schebesch, Ralf Stecking

Performance measures in discrete supervised classification
Ana Sousa Ferreira, Anabela Marques

Empirical comparison of recommendation strategies for legal documents on the web
Ruta Petraityte, Ansgar Scherp, Berthold Lausen

Multi-loss CNN architecture for image classification
Jian Piao, Mingzhe Jin

09.00 – 10.40 **SP1: Data mining techniques and classification methods in Social Sciences**
organized by A. Kalamatianou
Chair: Aglaia Kalamatianou — Room II

TV channels and predictive models: an analysis on social media
Paolo Mariani, Andrea Marletta, Mauro Mussini, Mariangela Zenga

Data mining techniques in autobiographical studies. Is there a chance?
Franca Crippa, Angela Tagini, Fulvia Mecatti

Requirements and competencies for labour market using conjoint analysis
Paolo Mariani, Andrea Marletta, Mauro Mussini, Mariangela Zenga

A data mining framework for Gender gap on academic progress
Aglaia Kalamatianou, Adele Marshall, Mariangela Zenga

Classification through graphical models: evidences from the EU-SILC data
Federica Nicolussi, Agnese Maria Di Brisco, Manuela Cazzaro

09.00 – 10.40 **CONTR2: Data science education**
Chair: Sofia Anastasiadou — Room III

Cross-disciplinary higher education of data science – beyond the computer science student
Evangelos Pournaras

The implications of network science in economic analysis
Éva Kuruczleki

Conception of measures of central tendency of primary school teachers
Evanthis Chatzivasileiou

09.00 – 10.40 **CONTR3: Classification and clustering in biological and medical research I**
Chair: George Menexes — Room IV

Quality of life profiles of colon cancer survivors: A three-step latent class analysis
Felix J. Clouth, Gijs Geleijnse, Lonneke van de Poll-Franse, Steffen Pauws, Jeroen Vermunt

Classifying functional groups of microorganisms with varying prevalence level using 16S rRNA
Rafal Kulakowski, Etienne Low-Decarie, Berthold Lausen

Identifying Chronic Obstructive Pulmonary Disease (COPD) phenotypes to predict treatment response
Vasilis Nikolaou, Sebastiano Massaro, Masoud Fakhimi, Lampros Stergioulas

The use of gene ontology to improve gene selection process for omics data analysis
Chadia Ed-driouch, Hassan Kafsaoui, Ahmed Moussa

10.40 – 11.10 **Coffee Break**

11.10 – 12.00 **President's Invited Lecture: Deciding what's what: classification from A to Z**
David Hand
Chair: Berthold Lausen — Plenary Hall

12.00 – 12.50 **Award Session**
Chair: Maurizio Vichi — Plenary Hall
IFCS 2019 Research Medal: David Hand
IFCS 2019 Chikio Hayashi Award: Sanjeena Dang, Abby Flynt, Brian Franczak, Cristina Tortora
IFCS 2019 Helga and Wolfgang Gaul Stiftung Award: Aghiles Salah
IFCS 2019 Student/Postdoctoral Fellow Paper Competition and Travel Award: Ali Mertcan Köse, Hanneke van der Hoef

13.00 – 14.00 **Mid-day Break**

Parallel Sessions

14.00 – 15.20 **SP2: Classification, visualisation and dimension reduction I**
dedicated to the late John Gower and organized by N. J le Roux
Chair: Sugnet Lubbe — Room I

Properties of individual differences scaling and its interpretation
Niël J le Roux, John Gower

Local and global relevance of features in multi-label classification
Trudie Sandrock

A multivariate ROC based classifier
Martin Kidd

Functional linear discriminant analysis for several functions and more than two groups
Sugnet Lubbe

14.00 – 15.20 **SP3: Advances in mixture modeling**
organized by S. Ingrassia
Chair: Salvatore Ingrassia — Room II

Variable selection in linear regression models with non-gaussian errors: a Bayesian solution
Giuliano Galimberti, Saverio Ranciati, Gabriele Soffritti

Finite mixtures of matrix-variate regressions with random covariates
Salvatore Daniele Tomarchio, Paul McNicholas, Antonio Punzo

Telescoping mixtures - Learning the number of components and data clusters in Bayesian mixture analysis,
Gertraud Malsiner-Walli, Sylvia Frühwirth - Schnatter, Bettina Grün

Finite mixture modeling and model-based clustering for directed weighted multilayer networks
Volodymyr Melnykov, Shuchismita Sarkar, Yana Melnykov

14.00 – 15.20 **SP4: Philosophy relevant to classification and data science**
organized by C. Hennig
Chair: Christian Hennig — Room III

The epistemology of nondistributive profiles
Patrick Allo Skype Presentation

Prediction without estimation: a case study in computer vision
Jérémy Grosman

Reconceptualizing null hypothesis testing
Jan Sprenger

14.00 – 15.20 **CONTR4: Official statistics**
Chair: Athanasios Thanopoulos — Room IV

Intertemporal exploratory analysis of Greek households in relation to information and communications technology (ICT) from official statistics
Stratos Moschidis, Athanasios Thanopoulos

Hierarchical clustering for anonymization of economic survey data
Kiyomi Shirakawa, Takayuki Ito

Improvement of training data based on pattern of reliability scores for overlapping classification
Yukako Toko, Mika Sato-Ilic, Shinya Iijima

14.00 – 15.20 **SP5: Data science education I (President's invited session)**
organized by B. Lausen and T. Chadjipadelis
Chair: Berthold Lausen — Room V

Progress of statistics and data science education in Japanese universities
Akimichi Takemura

Before Teaching Data Science, Let's First Understand How People Do It
Rebecca Nugent

15.20 – 16:10 **Plenary Invited: Modeling Networks and Network Populations via Graph Distances**
Sofia Olhede
Chair: Christian Hennig — Plenary Hall

16.10 – 17:00 **Plenary Invited: Principles for building your own machine learning methods: From theory to applications to practice**
Theodoros Evgeniou
Chair: Odysseas Moschidis — Plenary Hall

17.00 – 17.30 **Coffee Break**

Parallel Sessions

17.30 – 18.50 **CONTR5: Dimension reduction and clustering I**
Chair: Alfonso Iodice D'Enza — Room I

Simultaneous clustering and dimension reduction on multi-block data
Shuai Yuan, Katrijn Van Deun

Model-based hierarchical parsimonious clustering and dimensionality reduction
Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria

Active labeling using model-based classification
Cristina Tortora

Chunk-wise PCA with missings
Alfonso Iodice D'Enza, Angelos Markos, Francesco Palumbo

17.30 – 18.50 **CONTR6: Correspondence analysis I**
Chair: Giannoula Florou — Room II

Multidimensional data analysis of shopping records towards knowledge-based recommendation techniques
George Stalidis, Pantelis Kaplanoglou, Kostas Diamantaras

Principal Component Analysis to explore social attitudes towards the green infrastructure plan of Drama city
Vassiliki Kazana, Angelos Kazaklis, Dimitrios Raptis, Efthimia Chrisanthidou, Stella Kazakli, Nefeli Zagourgini

MCA's visualization techniques: an application to social data
Vasileios Ismyrlis, Efstratios Moschidis, Theodoros Tarnanidis

17.30 – 18.50 **SP6: Data analysis in finance**
organized by K. Jajuga
Chair: Krzysztof Jajuga — Room III

Sentiment and return distributions on the German stock market
Emile David Hövel, Matthias Gehrke

Risk management based on conditional extreme quantile risk measures on energy market
Grażyna Trzpiot, Alicja Ganczarek-Gamrot, Dominik Krężołek

Comparison of systemic risk in the banking sector and selected sectors of real economy – case of Poland
Katarzyna Kuziak, Krzysztof Piontek

Credit risk with credibility theory: a distribution-free estimator for probability of default, value at risk and expected shortfall
Anne Sumpf

17.30 – 18.50 **CONTR7: Algorithms for clustering and classification I**
Chair: Akinori Okada — Room IV

Flexible clustering
Andrzej Sokołowski, Małgorzata Markowska

A coefficient of determination for clusterwise linear regression with mixed-type covariates
Salvatore Ingrassia, Roberto Di Mari

Triplet clustering of one-mode two-way proximities
Akinori Okada, Satoru Yokoyama

17.30 – 18.50 **SP7: Data science education II (President's invited session)**
organized by B. Lausen and T. Chadjipadelis
Chair: Theodore Chadjipadelis — Room V

Societal responsibility of data scientists
Ursula Garczarek, Detlef Steuer

Data Science Education, Skills and Industry in Europe
Berthold Lausen, Alexander Partner, Stephen Lee, Henrik Nordmark, Mahdi Salhi, Christopher Saker

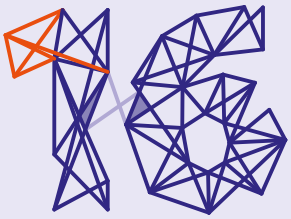
Discussion on Data Science Education
moderated by Berthold Lausen, Theodore Chadjipadelis

18.50 – 19:40 **Plenary Invited: Correspondence analysis: Jack of all trades, Master of one**
Michael Greenacre
Chair: Patrick Groenen — Plenary Hall

19:40 - 20:15 **Data Science, Elections and Government**

Theodoros Livanios, *Deputy Minister for Local Government and Elections, Greece*

Plenary Hall



Wednesday, 28th August 2019

08.00 – 09.00 Registration

Parallel Sessions

09.00 – 10.40 **SP8: Classification methods in economics and business**
organized by J. Pociecha
Chair: Józef Pociecha — Room I

Analysis of statistical tests indications in assessing data conformity to Benford's Law in fraud detection
Józef Pociecha, Mateusz Baryła

Conditional extreme quantile risk measures on metals market
Dominik Krężołek, Grażyna Trzpiot

Fuzzy clustering with skew components, with applications in Economics and Business
Francesca Greselin, Luis Angel Garcia-Escudero, Agustín Mayo-Iscar

Distance measurement and clustering when fuzzy numbers are used. Survey of selected problems and procedures
Jozef Dziechciarz, Marta Dziechciarz Duda

The impact of the publication of short selling positions on German stock returns
Matthias Gehrke, Jannis Kepper

09.00 – 10.40 **CONTR8: Topic models, document clustering and classification I**
Chair: Boris Mirkin — Room II

Japanese women's attitudes towards childrearing: text analysis and multidimensional scaling
Kunihiro Kimura

Using domain taxonomy for computational generalization
Boris Mirkin, Dmitry Frolov, Susana Nascimento, Trevor Fenner

Detection of topics and time series variation in consumer web communication data
Atsuho Nakayama

Making product recommendations based on latent topics: an analysis of online purchase data with topic models
Johanna Fischer

Quantitative analysis of phonological structure used in dialects in Osamu Dazai's works
Naoko Oshiro, Sayaka Irie, Mingzhe Jin

09.00 – 10.40 **CONTR9: Supervised learning and applications II**
Chair: Katrijn Van Deun — Room III

Isotonic boosting procedures for classification
Miguel Fernández, David Conde, Cristina Rueda, Bonifacio Salvador

Development of indices for the regional comparative analysis of musical compositions, focusing on rhythm
Akihiro Kawase, Mitsuru Tamatani

View selection through meta-learning
Wouter van Loon, Marjolein Fokkema, Botond Szabo, Mark de Rooij

The δ-machine: Classification based on distances towards prototypes
Beibei Yuan, Willem Heiser, Mark de Rooij

09.00 – 10.40 **SP9: Classification, visualisation and dimension reduction II**
organized by N. J. le Roux
Chair: Niël J le Roux — Room IV

Tree-base ensemble methods for classification
Daniel Uys

Unsupervised feature selection and big data
Renato Cordeiro De Amorim

A simulation study for the identification of missing data mechanisms using visualisations
Johané Nienkemper-Swanepoel, Niël J le Roux, Sugnet Lubbe

Visualising Multivariate Data in a Principal Surface Biplot
Raeesa Ganey, Sugnet Lubbe

Using separate sampling to understand mobile phone security compliance
Rénette Blignaut, Isabella Venter, Humphrey Brydon

09.00 – 10.40 **SP10: Clustering categorical and mixed-type data**
organized by C. Tortora
Chair: Cristina Tortora — Room V

Model based clustering through copulas: parsimonious models for mixed mode data
Dimitris Karlis, Ioannis Kosmidis, Fotini Panagou

Clustering ranked data using copula
Marta Nai Ruscone

Linking different kinds of omics data through a model-based clustering approach
Vincent Vandewalle, Camille Ternynck, Guillemette Marot

A probabilistic distance algorithm for nominal data
Francesco Palumbo, Mario Migliaccio, Cristina Tortora

Stability of joint dimension reduction and clustering
Michel van de Velden, Angelos Markos, Alfonso Iodice D'Enza

10.40 – 11.10 **Coffee Break**

Parallel Sessions

11.10 – 12.50 **CONTR10: Algorithms for clustering and classification II**
Chair: Francesco Palumbo — Room I

Hierarchical clustering through a penalized within-cluster sum-of-squares criterion
Patrick Groenen, Yoshikazu Terada, Mariko Takagishi

PerioClust: a new Hierarchical Agglomerative Clustering method including temporal ordering constraints
Lise Bellanger, Arthur Coulon, Philippe Husi

Iterated dissimilarities and some applications
François Bavaud

Constrained three-way clustering around latent variables approach
Véronique Cariou, Tom Wilderjans

Clustering binary data by application of combinatorial optimization heuristics
Javier Trejos, Luis Amaya, Alejandra Jiménez, Alex Murillo, Eduardo Piza, Mario Villalobos

11.10 – 12.50 **CONTR11: Classification and clustering in biological and medical research II**
Chair: Fionn Murtagh — Room II

Testing for equation of distance-based regressions to see whether two groups form a species
Christian Hennig, Bernhard Hausdorf

Mental health: analytical focus and contextualization for deriving mental capital
Fionn Murtagh

A deep learning analytics to detect prognosis of HCC
Taerim Lee

Analysis of the regional difference of number of patients with blood coagulation disorders in Japan
Shinobu Tatsunami, Kagehiro Amano, Akira Shirahata, Masashi Taki

11.10 – 12.50 **CONTR12: Data science in economics and business I**
Chair: Tadashi Imaizumi — Room III

Analysis of the Power Balance of the companies of the "keirersu" with the Asymmetric MDS
Tadashi Imaizumi

A fast-electric vehicle planner using clustering
Jaël Champagne Gareau, Vladimir Makarenkov, Éric Beaudry

The technology innovation and the critical raw material stock
Beatrix Margit Varga, Kitti Fodor

Knowledge graph mining and affinity analysis for product recommendation on online-marketplace platforms
Siti Nur Muninggar, Reza Aditya Permadi, Simon Simbolon, Verra Mukty, Putri Wikie Novianti

Pension expenditure modelling and classification analysis
Kimon Ntotsis, Marianna Papamichail, Peter Hatzopoulos, Alex Karagrigoriou

11.10 – 12.50 **SP11: Supervised classification with imprecise labels and complex data**
organized by A. Montanari
Chair: Angela Montanari — Room IV

Estimation of classification rules from partially classified data
Geoffrey McLachlan

Classification with imperfect training labels
Timothy Cannings, Yingying Fan, Richard Samworth

Classification with unknown class conditional label noise on non-compact feature spaces
Henry Reeve, Ata Kaban

Supervised classification of long or unbalanced datasets
Laura Anderlucci, Roberta Falcone, Angela Montanari

11.10 – 12.50 **CONTR13: Modeling of psychological processes and clustering in educational research**
Chair: Eva Ceulemans — Room V

Kernel change point detection on the running statistics: A flexible, comprehensive and user-friendly tool
Eva Ceulemans, Jedelyn Cabrieto, Kristof Meers, Janne Adolf, Peter Kuppens, Francis Tuerlinckx

School motivation profiles of students in secondary education
Matthijs Warrens, Denise Blom

Probing the nature of psychological constructs with Taxometrics and Latent Class Analysis: The case of children's mental models
Dimitrios Stamovlasis, Julie Vaiopoulou, George Papageorgiou

On the use and reporting of cluster analysis in educational research: A systematic review
Hanneke van der Hoef, Matthijs Warrens, Marieke Timmerman

11.10 – 14.00 **Poster Session** Poster Area
Chairs: Sofia Anastasiadou, Odysseas Moschidis

The relationship of the apolipoprotein E genotype gene to the Alzheimer's Disease: A meta-analysis
Sofia Anastasiadou

Bayesian analysis for chromosomal interactions in hi-c data using hidden Markov random field model
Osuntoki Intunu, Andrew Harrison, Hongsheng Dai, Yanchun Bao, Nicolae Zabet

New financial instruments: Pollution emission rights and their trading on the stock exchange
Argiro Dimitoglou

Econometric assessment of the relation between the situation of youth on the labour market and macroeconomic situation among the EU countries
Beata Bal-Domańska, Elżbieta Sobczak

Comparison of pattering methods: Clustering of variables, Implicative Statistical Analysis and Correspondence Analysis
Sofia Anastasiadou

Framing coworking spaces digital marketing strategy via social media analytics
Dimitrios Vagianos, Nikos Koutsoupias

Sales performance measure: A systematic review and typology of research studies
Tor Korneliussen, Per Ivar Seljeseth, Michael Greenacre

Document clustering via multiple correspondence, term and metadata analysis in R
Nikos Koutsoupias, Kyriakos Mikelis

Comparison of multivariate methods in group/cluster identification: PCA vs Discriminant Analysis and K-Means clustering
Sofia Anastasiadou

Asymptotic cumulants of the minimum phi-divergence estimator for categorical data under possible model misspecification
Haruhiko Ogasawara

Multidimensional data analysis in perception of European Union by different generations
Agnieszka Stanimir

13.00 – 14.00 **Mid-day Break**

14.00 – 14.50 **Presidential Address: Predictive ensemble methods for event time data**
Berthold Lausen
Chair: Angela Montanari

14.50 – 15.40 **Plenary Invited: On the consistency of supervised learning with missing values**
Julie Josse
Chair: Angelos Markos Plenary Hall

15.40 – 16.10 **Coffee Break**

Parallel Sessions

16.10 – 17.10 **SP12: New developments in clustering and scaling data**
organized by J. F. Vera & E. Boj del Val
Chair: Jose Luis Vicente-Villardón Room I

A cellwise trimming approach to Cluster Analysis
Luis Angel Garcia-Escudero, Diego Rivera-García, Joaquín Ortega, Agustín Mayo-Iscar

Redundancy analysis for categorical data based on logistic regressions
Jose Luis Vicente-Villardón, Laura Vicente-Gonzalez

A log-ratio approach to cluster analysis of count data when the total is irrelevant
Marc Comas-Cufi, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, Javier Palarea-Albaladejo

16.10 – 17.10 **SP13: Developments of data analysis in Greece**
organized by I. Papadimitriou and T. Chadjipadelis
Chair: Giannoula Florou — Room II

Doing research and teaching data analysis in Greek higher education
Iannis Papadimitriou, Vicky Bouranta

Data Analysis Bulletin
Dimitris Karapistolis, Marina Sotiroglou

The Past, the Presence and the Future (round table)
Ilias Athanasiadis, Giannoula Florou, Georgia Panagiotidou

16.10 – 17.10 **CONTR14: Symbolic data**
Chair: Paula Brito — Room III

Clustering and classification of interval time series
Ann Maharaj, Paula Brito, Paulo Teles

Multiple-valued symbolic data clustering using regression mixtures of Dirichlet distributions
José Dias

Visualization of heterogeneity in exploratory meta-analysis
Masahiro Mizuta

16.10 – 17.10 **CONTR15: Multivariate visualization**
Chair: Adalbert Wilhelm — Room IV

QVisVis: Framework and R toolkit for Exploring, Evaluating, and Comparing Visualizations
Stephen L. France, Ulas Akkucuk

Visual exploration for feature extraction and feature engineering
Adalbert Wilhelm

Multivariable analysis on the use of social media & web 2.0/3.0. Modeling & clustering of users
Evangelia Nikolaou Markaki, Theodore Chadjipadelis

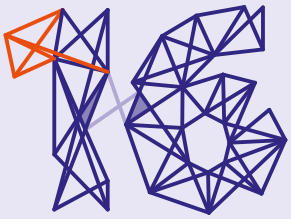
16.10 – 17.10 **CONTR16: Data science in economics and business II**
Chair: George Stalidis — Room V

Probabilistic collaborative representation learning
Aghiles Salah, Hady Lauw

User profiling for a better search strategy in e-commerce website
Putri Wikie Novianti, Fatia Kusuma Dewi

Comparison of the sharing economy participants' motivation
Roland Szilágyi, Levente Lengyel

17.10 – 18.00	**Plenary Invited: Model-based clustering without parametric assumptions** **David Hunter** **Chair:** Akinori Okada Plenary Hall
18.00 – 19.45	IFCS Council Meeting
19.10	Bus City Tour
20.30	Conference Dinner



Thursday, 29th August 2019

08.00 – 09.00 Registration

Parallel Sessions

09.00 – 10.40 **SP14: Clustering for spatio-temporal data and its visualization**
organized by K. Kurihara
Chair: Koji Kurihara — Room I

Classification of suicidal execution area in Japan by areal statistics of committed suicide
Takafumi Kubota

Visualization and provision method of meteorological data for Energy Management System
Yoshiro Yamamoto

Spatial perception for structured and unstructured data in topological data analysis
Yoshitake Kitanishi, Fumio Ishioka, Masaya Iizuka, Koji Kurihara

Dimensional reduction clustering with modified outcome method
Kensuke Tanioka, Hiroshi Yadohisa

09.00 – 10.40 **SP15: Emerging methodologies in economics and finance**
organized by T. Papadimitriou and P. Gogas
Chair: Theophilos Papadimitriou — Room II

Forecasting transportation demand for the U.S. market
Vasilios Plakandaras, Theophilos Papadimitriou, Periklis Gogas

Money neutrality, monetary aggregates and machine learning
Emmanouil Sofianos, Theophilos Papadimitriou, Periklis Gogas

Forecasting S&P 500 spikes: an SVM approach
Athanasios-Fotios Athanasiou, Theophilos Papadimitriou, Periklis Gogas

Assessing the resilience of the U.S. banking system
Anna Agrapetidou, Theophilos Papadimitriou, Periklis Gogas

09.00 – 10.40 **SP16: Analysis of European parliament elections**
organized by T. Exadaktylos, T. Chadjipadelis
Chair: Theodore Chadjipadelis — Room III

What was really the case? Party competition in Europe at the occasion of the 2019 European Parliament Elections
Theodore Chadjipadelis, Eftichia Teperoglou

First-time voter in Greece: Views and attitudes of youth on Europe and Democracy
Georgia Panagiotidou, Theodore Chadjipadelis

Developing a model for the analysis of the political programmes
Theodore Chadjipadelis, Panagiotis Paschalidis

How the undecided voters decide?
George Siakas

09.00 – 10.40 **CONTR17: Topic models, document clustering and classification II**
Chair: Kunihiro Kimura — Room IV

Improving the performance of Japanese authorship attribution with phonetic related information
Hao Sun, Mingzhe Jin

Double helix multi-stage text classification model to enhance chat user experience in e-commerce website
Fiqry Revadiansyah, Abdullah Ghifari, Rya Meyvriska

Latent dimensions of the museum experience: the role of the online reviews
Melisa Diaz, Anna Calissano

A corpus-based approach to explore the stylistic peculiarity of Kouji Uno's postwar works
Xueqin Liu, Mingzhe Jin

10.40 – 11.10 Coffee Break

Parallel Sessions

11.10 – 12.50 CONTR18: Network analysis and applications
Chair: Vladimir Batagelj — Room I

The analyses of the WoS data on network clustering
Anuška Ferligoj, Vladimir Batagelj, Patrick Doreian

Approximate core-and-shell supercluster in statics and dynamics
Boris Mirkin

Trust your data or not - Standard remains Standard (QP); implications for robust clustering in social networks
Immanuel Bomze, Michael Kahr, Markus Leitner

Classifying users through keystroke dynamics
Ioannis Tsimperidis, George Peikos, Avi Arampatzis

11.10 – 12.50 SP17 Neutral Benchmarking Studies of Clustering
organized by the IFCS Cluster Benchmarking Task Force
Chair: Iven Van Mechelen, **Moderator:** Christian Hennig — Room II

Introduction
Matthijs Warrens

K-means, spectral clustering, or DBSCAN: a benchmarking study
Irene Cho, Nivedha Murugesan, Cristina Tortora*

Benchmarking minimax linkage
Xiao Hui Tai, Kayla Frisoli

Benchmarking in cluster analysis for mixed-type data
Madhumita Roy, Jarrett Jimeno, Cristina Tortora*

Comparison of dimensionality reduction and cluster analysis methods for high dimensional datasets
Jingfei Gong, Yuwen Luo, Cristina Tortora*

Evaluation of text clustering methods and their dataspace embeddings: an exploration
Alain Lelu, Martine Cadot

*video-recorded presentation

11.10 – 12.50 SP18: Advances in classification analysis for complex data – compositional and symbolic approaches
organized by N. Kejžar, S. Korenjak-Černe, A. Srakar
Chair: Simona Korenjak-Černe — Room III

Clustering transformed compositional data using K-means, with applications in gene expression and bicyle sharing system data
Antoine Godichon-Baggioni, Cathy Maugis-Rabusseau, Andrea Rau

Entrepreneurial regimes classification: a symbolic polygonal clustering approach
Andrej Srakar, Marilena Vecco

Distances and discriminant analysis for microbial communities composition to classify inflammatory bowel diseases
Glòria Mateu-Figueras, Pepus Daunis-i-Estadella, Mireia López-Siles, Josep Antoni Martín-Fernández

Symbolic data analysis of gender-age-cause-specific mortality in European countries
Filipe Afonso, Aleša Lotrič Dolinar, Simona Korenjak-Černe, Edwin Diday

---

11.10 – 12.50 **SP19: Clustering, classification and data analysis via mixture models**
organized by P. McNicholas
Chair: Brian Franczak — Room IV

Clustering multivariate count data using a family of multivariate Poisson log-normal distributions
Sanjeena Dang

Growth mixture modeling with measurement selection
Abby Flynt, Nema Dean

On the use of multiple scaled distributions for outlier detection and model-based learning
Brian Franczak, Antonio Punzo, Cristina Tortora

Skewed distributions or transformations? Accounting for skewness in cluster analysis
Michael Gallaugher, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu

13.00 – 14.00 **Mid-day Break**

14.00 – 14.50 **Plenary Invited: Clustering in networks**
Vladimir Batagelj
Chair: Angela Montanari — Plenary Hall

Parallel Session

14.50 – 15.50 **CONTR19: Dimension reduction and clustering II**
Chair: Vartan Choulakian — Room I

Clustering of variables using CDPCA
Adelaide Freitas

A study of the variable outlyingness ranking that is obtained using different loading similarity coefficients
Sopiko Gvaladze, Kim De Roover, Francis Tuerlinckx, Eva Ceulemans

Some properties of coherent clusters of rank data
Vartan Choulakian

14.50 – 15.50 **CONTR20: Classification and regression trees** Skype Presentations
Chair: Iven van Mechelen — Room II

C443: A methodology to see a forest for the trees
Iven Van Mechelen, Aniek Sies

Assessing how feature selection and hyper-parameters influence optimal trees ensemble and random projection
Nosheen Faiz, Naz Gul, Metodi Metodiev, Andrew Harrison, Zardad Khan, Berthold Lausen

Residual diagnostics for model-based trees for ordinal responses
Rosaria Simone, Carmela Cappelli, Francesca Di Iorio

14.50 – 15.50 **CONTR21: Functional data**
Chair: Tomasz Górecki — Room III

Measuring and testing mutual dependence for functional data
Tomasz Górecki, Mirosław Krzyśko, Waldemar Wołyński

A co-clustering method for multivariate functional curves
Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Chèze, Pauline Martin

One-way repeated measures ANOVA for functional data
Łukasz Smaga

14.50 – 15.50 **SP20: Statistical theory of cluster analysis**
organized by D. Hunter
Chair: David Hunter — Room IV

Hidden Markov models for continuous multivariate data with missing responses
Fulvia Pennoni

Mixtures of cluster-weighted models with latent factor analyzer structure
Sanjeena Dang

Specification of basis spacing for process convolution Gaussian process models
Herbert Lee, Waley Liang

14.50 – 15.50 **CONTR23: Classification and clustering in biological and medical research III**
Chair: Taerim Lee — Room V

Comparison of hierarchical clustering methods for binary data from SSR and ISSR molecular markers
Emmanouil Pratsinakis, Lefkothea Karapetsi, Symela Ntoanidou, Angelos Markos, Panagiotis Madesis, Ilias Eleftherohorinos, George Menexes

Inspecting smoking addiction of youth in Turkey through a latent class analysis
Ali Mertcan Köse, Elif Çoker

Data analysis on the annual use of the new deferasirox formulation in pediatric thalassemia patients
Alkistis Adramerina, Aikaterini Teli, Symeon Symeonidis, Nikoleta Printza, Antonios Papastergiopoulos, Labib Tarazi, Emmanouil Chatzipadelis, Marina Economou

15.50 – 16.20 **Coffee Break**

Parallel Sessions

16.20 – 17.40 **SP21: Crossroads of statistical learning and psychometrics**
organized by M. de Rooij
Chair: Mark de Rooij — Room I

Recursive partitioning of longitudinal and growth curve models
Marjolein Fokkema

Bayesian regularization in probabilistic PCA with sparse weights matrix
Davide Vidotto

Gaussian process panel modeling – statistical learning inspired analysis of longitudinal panel data
Julian Karch, Andreas Brandmaier, Manuel Voelkle

Finding the hidden link: Sparse common component analysis
Katrijn Van Deun

16.20 – 17.40 **CONTR22: Correspondence analysis II**
Chair: Michel van de Velden — Room II

Variants of three-way correspondence analysis: An R package
Rosaria Lombardo, Michel van de Velden, Eric Beh

Another view of Correspondence Analysis through Design and Projection matrices and General Linear Models
George Menexes, Angelos Markos, Emmanouil Pratsinakis

Implicative and conjugative variables in the context of Correspondence Analysis
Odysseas Moschidis, Angelos Markos

Combined use of Correspondence Analysis and Ordinary kriging to display "supplementary" values of quantitative variables onto the factorial planes
Georgios Menexes, Thomas Koutsos

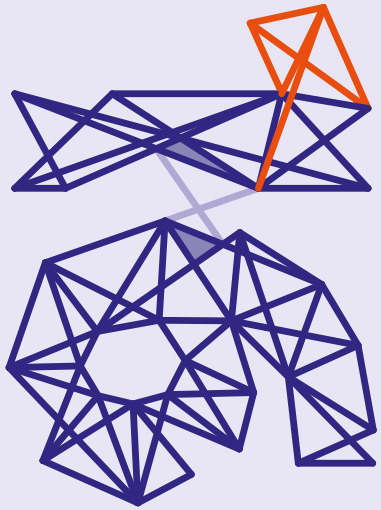
16.20 – 17.40 **CONTR24: Model-based clustering**
Chair: Geoffrey McLachlan — Room III

On missing label patterns in semi-supervised learning
Daniel Ahfock, Geoffrey McLachlan

Bayesian nonparametric mixture modeling for ordinal regression
Athanasios Kottas, Maria DeYoreo

Assessment of recent social attitudes in Japan: a latent class item response theory model for web survey data
Fulvia Pennoni, Miki Nakai

17.40 – 18.10 Benchmarking Challenge Award
Presentation of IFCS-2021
Closing Ceremony — Plenary Hall



KEYNOTE LECTURES

Clustering in networks

Vladimir Batagelj

Abstract Real life networks are usually created by some processes adding / removing nodes or links and changing their properties. An increased activity in some part of a network often increases a local density of nodes / links and intensity of properties in that part. In this talk we will consider two problems: identification of important units / subnetworks and determining a (complete) clustering of a given network. To identify important subnetworks we usually define a measure (property of nodes or weight on links) expressing our goal / question combining local structural information (indexes, fragments, motiffs, graphlets) with available variables. The important subnetworks are determined using procedures such as cuts and islands. Many approaches exsist to the clustering in networks. We will limit our attention to:

- clustering with relational constraint: clusters are connected subgraphs of selected type containing similar units;
- blockmodeling: the reduced network obtained by shrinking the clusters provides a good description of the overall structure of the network.

For details see:

1 Doreian P, Batagelj V, Ferligoj A: Generalized blockmodeling. Cambridge UP, 2004.

2 Batagelj V, Doreian P, Ferligoj A, Kejˇzar N: Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley, 2014.

3 Doreian P, Batagelj V, Ferligoj A: Advances in Network Clustering and Blockmodeling. Wiley, 2019.

Vladimir Batagelj
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
University of Primorska, Koper, Slovenia
National Research University Higher School of Economics, Moscow, Russia
e-mail: vladimir.batagelj@fmf.uni-lj.si

Principles for Building your Own Machine Learning Methods: From Theory to Applications to Practice

Theodoros Evgeniou

Abstract Machine Learning has greatly matured as a field the past 20-30 years or so. While it is now widely spread and used by researchers and business, the majority of the tools are "out of the box" ones, which have been developed for largely "general purpose problems" in some sense. The focus of this presentation will be to go back to some of the key theoretical principles of machine learning methods, and based on the "basics", discuss examples of developing custom-based machine learning methods that may fit the needs of a specific problem/application. A few such "custom built" methods will be presented, with example applications ranging from understanding choices people make (e.g., purchase ones) to investing in stocks. Moreover, some part of the presentation will be spent on placing all recent advances in Machine Learning and AI in some broader context – of how they may, or may not fit in practice, whether it is in research, business, and more broadly in society. Research questions – beyond the "mathematics of machine learning" - will be discussed.

Theodoros Evgeniou
Professor of Decision Sciences and Technology Management at INSEAD in Fontainebleau France and an Academic Director of INSEAD eLab, France, e-mail: theodoros.evgeniou@insead.edu

Correspondence analysis: Jack of all trades, Master of one

Michael Greenacre

Abstract Correspondence analysis (CA) as a method of multivariate data visualization has an almost 60-year history and I have been a part of it for 46 years since my doctoral studies in Prof. Jean-Paul Benzécri's laboratory in Paris in the years 1973-1975. In this talk, I reflect on (i) how this method has developed theoretically and matured over the last decades, (ii) what distinguishes it from other approaches to dimension reduction, and (iii) why this method is an important and essential addition to the applied multivariate toolbox.

These days CA has found use in almost every field of multivariate research. Published applications as well as methodological papers on CA have been rising exponentially, especially in the fields of biology and ecology. CA also has intimate connections with other multivariate areas, notably compositional data analysis, discriminant analysis, analysis of variance and principal component analysis. Because CA is applicable to the most basic of data types, namely categorical data, it can be applied to almost any multivariate data set, thanks to ingenious ways of recoding data to categorical scales. But the prime example, where all its properties make perfect sense, is that of bivariate categorical data in the form of a contingency table, and its generalization to multivariate categorical data. This talk will include several emblematic applications of CA as well as personal musings on the future of what can be called "the joy of CA".

Keywords Multivariate analysis; dimension reduction; categorical data; contingency tables

Michael Greenacre
Professor of Statistics at the Universitat Pompeu Fabra, Barcelona, Spain, e-mail: michael.greenacre@upf.edu

Deciding what's what: classification from A to Z

David J. Hand

Abstract Our lives are constructed in terms of classes: fatal or non-fatal diseases, spoken words indicating affirmative or negative, urgent and not-so-urgent emails, trains which will get us there on time versus those which will not, and so on endlessly. But these classifications themselves come in two classes: those which are forced upon us, and those which we force upon nature. I explore this difference, examining classification ideas in many different domains, showing how the discovery of classes has led to scientific breakthroughs and how the imposition of classes has enabled progress. I also examine formal methods for producing classifications, and see how these have evolved since a moratorium on the development of new mathematical methods was proposed (but generally ignored) in the 1980s.

David J. Hand
Emeritus Professor of Mathematics and Senior Research Investigator at Imperial College, London, United Kingdom,
e-mail: d.j.hand@imperial.ac.uk

Model-Based Clustering without Parametric Assumptions

David Hunter

Abstract This talk discusses finite mixture models in which the component distributions are notassumed to come from any particular parametric family. We begin with a discussion of the essential question of parameter identifiability and introduce an EMlike framework often used to fit these models. We extend these ideas to the multivariate case, which is actually easier in some sense than the univariate case, and introduce the important assumption of conditional independence. We show how to construct an EM-like algorithm, based on a majorization-minimization idea, with desirable theoretical properties. Finally, we extend the multivariate model so that conditional independence need not be assumed. This extension uses a welldeveloped technique known as independent component analysis (ICA) to create a hybrid estimation algorithm. We illustrate this new methodology using applications in unsupervised learning and image processing, and we discuss what is and is not known about the theoretical properties of the model and the algorithm.

David Hunter
Professor at Penn State Department of Statistics, United States, e-mail: dhunter@stat.psu.edu

On the consistency of supervised learning with missing values

Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux

Abstract In many application settings, the data have missing features which make data analysis challenging. An abundant literature addresses missing data in an inferential framework: estimating parameters and their variance from incomplete tables. Here, we consider supervised-learning settings: predicting a target when missing values appear in both training and testing data. We show the consistency of two approaches in prediction. A striking result is that the widely-used method of imputing with the mean prior to learning is consistent when missing values are not informative. This contrasts with inferential settings where mean imputation is pointed at for distorting the distribution of the data. That such a simple approach can be consistent is important in practice. We also show that a predictor suited for complete observations can predict optimally on incomplete data, through multiple imputation. We analyze further decision trees. These can naturally tackle empirical risk minimization with missing values, due to their ability to handle the halfdiscrete nature of incomplete variables. After comparing theoretically and empirically different missing values strategies in trees, we recommend using the ``missing incorporated in attribute'' method as it can handle both non-informative and informative missing values.

Julie Josse
(CMAP, XPOP), Professor of Statistics at Ecole Polytechnique, France e-mail: julie.josse@polytechnique.edu

Nicolas Prost
(CMAP, XPOP, PARIETAL), Ecole Polytechnique, France

Erwan Scornet
(X), Ecole Polytechnique, France

Gaël Varoquaux
(PARIETAL), Ecole Polytechnique, France

Recent Developments in DOE for Agricultural Research

Andy Mauromoustakos, and Bradley Jones

Abstract In this talk we will review some of the restrictions and limitations of the classical textbook designs. We will focus on the popular variations of Split-Plot experiments and in the screening designs. We will then demonstrate advantages of the modern experimental designs reviewing computer-generated designs with different optimality criteria. The examples will demonstrate how using these new tools the AG experimenter can gain valuable insight as to the choice and the possible benefits before conducting the actual research. We will also show examples situations where the newer designs outperform the classical designs by leading to correct answers with vast fewer resources.

Andy Mauromoustakos
Professor AG STAT LAB, University of Arkansas, United States, e-mail: amauro@uark.edu

Bradley Jones
Distinguished Research Fellow, JMP Division/SAS, United States

Modeling Networks and Network Populations via Graph Distances

Sofia Olhede

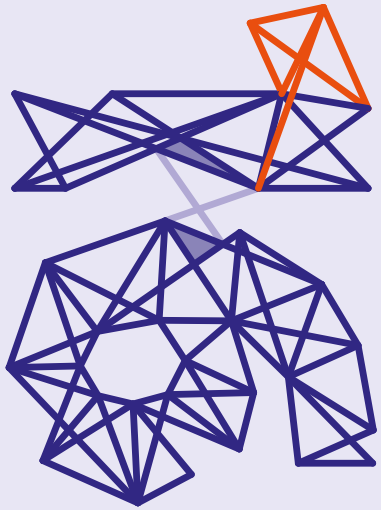
Abstract Networks have become a key data analysis tool. They are a simple method of characterising dependence between nodes or actors. Understanding the difference between two networks is also challenging unless they share nodes and are of the same size. We shall discuss how we may compare networks and also consider the regime where more than one network is observed.

We shall also discuss how to parametrize a distribution on labelled graphs in terms of a Frechét mean graph (which depends on a user-specified choice of metric or graph distance) and a parameter that controls the concentration of this distribution about its mean. Entropy is the natural parameter for such control, varying from a point mass concentrated on the Frechét mean itself to a uniform distribution over all graphs on a given vertex set.

Networks present many new statistical challenges. We shall discuss how to resolve these challenges respecting the non-Euclidean nature of network observations.

Keywords Statistical network analysis; Network variability; Graph metrics; Random graphs

Sofia Olhede
Professor of Statistics at University College London, Director of UCL's Centre for Data Science,
Honorary Professor of Computer Science, Senior Research Associate of Mathematics at University College London Department of Statistical Science, University College London, United Kingdom,
e-mail: sofia.olhede@epfl.ch=



ORALS

Multiclass posterior probability support vector machines for big data

Pedro Duarte Silva

Abstract Support Vector Machines (SVMs) were originally designed to handle two-class classification problems, and have quickly established themselves as one of the most accurate machine learning algorithms in terms of class prediction. However, this success did not translate to the problem of estimating posterior probabilities of class membership. In fact, early proposals to tackle this problem were empirically found to be often inferior to classical statistical methods for binary regression, and later theoretical results have shown that classical SVMs do not do not carry any information about the *a posterior* membership probabilities other than the predicted class membership. However, sequences of nonstandard SVMs with weighted loss functions (varying the weights for each SVM in the sequence) can recover posterior probabilities consistently and with competing performance in two-class real world applications. However, known extensions of this approach to problems with more than two classes lead to sequences of nonstandard non-convex SVM optimization problems that are computationally hard to solve, and can only be applied to small or moderate size problems.

In this talk we will revise these models, and propose revised model formulations and improved training algorithms that alleviate their computational burden, and make reliable multiclass posterior probabilities SVMs practical for bigger data problem.

Keywords support vector machines; categorical regression; training algorithms;

Pedro Duarte Silva
Católica Porto Business School and CEGE, Portugal, e-mail: psilva@porto.ucp.pt

Improving credit client classification by deep neural networks?

Klaus Bruno Schebesch, and Ralf Stecking

Abstract Estimating credit client default from past data is an important task the financial industry poses to statistical modeling. So far, there exists a broad spectrum of statistical methods, including Logistic Regression, the more recently developed Support Vector Machines and various Deep Learning models. Especially deep neural networks are by now frequently used for classification tasks. However, we do not find many applications for the task of credit client classification yet.

As we expect that future credit client classifiers may use much more complicated data types and bigger data volumes, we propose to analyze the suitability of deep neural networks which by their nature can easily process such input data. They can bypass the problem of vanishing backpropagation error signals even if many hidden layers of neurons are employed. Furthermore, efficient dropout layers find stable models even from grossly over-specified initial model architectures.

In our work, we use two different credit data sets of vastly different sizes in order to examine the classification performance of several classification methods. We found out that deep neural networks produce superior AUC on our test sets and likewise for a set of alternative optimizers, leading to the conclusion that Deep Learning may be successfully used to find still more accurate credit client classification models.

Keywords deep neural networks; credit client classification; method comparison

Klaus B. Schebesch
Vasile Goldis Western University Arad, Romania, e-mail: kbschebesch@uvvg.ro

Ralf Stecking
Carl von Ossietzky University Oldenburg, Germany, e-mail: ralf.w.stecking@uol.de

Performance measures in discrete supervised classification

Ana Sousa Ferreira, and Anabela Marques

Abstract The evaluation of results in Cluster Analysis frequently appears in the literature, and a variety of evaluation measures have been proposed. On the contrary, in supervised classification, particularly in the discrete case, the subject of results evaluation is relatively rare in the literature of the area and a part of the measures that have been proposed by some classification researchers are based on many of the measures used in Cluster Analysis. This is the motto for the present study. The evaluation of the performance of any model of supervised classification is, generally, based in the number of cases correctly and incorrectly predicted by the model. However, these measures can lead to a misleading evaluation when data is not balanced. More recently, another type of measures had been studied as coefficients of association or agreement, the Kappa statistics, the Huberty index, Mutual Information or even ROC curves. Exploratory studies have been made to understand the relationship between each measure and data characteristics, namely, samples size, balance and classes' separation. For this purpose, we resort to real and simulated data and use a generalization of the Tobit regression model on the performance of the models.

Keywords balanced classes; class separability; performance measures; supervised classification

Ana Sousa Ferreira
Universidade de Lisboa and Business Research Unit (BRU-IUL) Lisboa, Portugal, e-mail: asferreira@psicologia.ulisboa.pt

Anabela Marques
ESTBarreiro, Setúbal Polytechnic, Portugal and CIIAS-ESS, e-mail: anabela.marques@estbarreiro.ips.pt

Empirical comparison of recommendation strategies for legal documents on the web

Ruta Petraityte, Ansgar Scherp, and Berthold Lausen

Abstract When it comes to making legal document recommendations on the web, often one of the easiest strategies that tend to be applied by companies is to provide topic-based recommendations. Then, either Top-k most popular or most recently published documents from the same topic are shown to the user. While these are viable techniques - recommendations based on broad topics might not be as relevant or accurate.

We initially explore the potential of a well-established method in similar document retrieval, which is Top-k recommender using the classical term-weighting method TF-IDF (term frequency–inverse document frequency) with cosine similarity. The initial approach was chosen to serve as a baseline for further work, due to its stability and ability to identify the more important and rarer terms in a document, thus helping us identify more accurate similar document to what a user is reading at the time. The recommendations derived from TF-IDF based technique were then used to compare with the two different variations of the topic-based recommendations: recency and popularity based. User engagement (clicks) were then used in evaluating which recommendation system performed best.

In the period of 30 days, a total of 8,887 registered and 156,863 unregistered user clicks were gathered, where the TF-IDF method outperformed both topic recency and popularity based methods by receiving, on average, more than 120% and 100% of registered user clicks, and more than 130% and 380% of unregistered user clicks, respectively. The results have shown us the potential of using a classical Information Retrieval technique, which now will be used as a baseline for exploring more advanced methods.

For further work, Lasso will be explored as a dimensionality reduction technique to help both increase document relevancy and decrease the computational time it requires to find Top-k most similar documents. Additionally, two extensions of TF-IDF will be used, namely CF-IDF (concept, rather than term-weighting method) and HCF-IDF (a novel hierarchical, concept frequency driven variation of TF-IDF), to see whether the results can be improved upon by using semantic concept frequency, rather than term frequency as a document representation. The different techniques will then be evaluated by running a live online test, where different variations of recommendations will be presented to users to see which performs best.

Keywords recommendation system; information-retrieval; term frequency–inverse document frequency

Ruta Petraityte
Mondaq, Knowledge Gateway, United Kingdom, email: Ruta.Petra@Mondaq.com

Ansgar Scherp
University of Essex, United Kingdom, email: ansgar.scherp@essex.ac.uk

Berthold Lausen
University of Essex, United Kingdom, email: blausen@essex.ac.uk

Multi-loss CNN architecture for image classification

Jian Piao, and Mingzhe Jin

Abstract Deep convolutional neural networks have been widely used in image classification. This study proposes a novel multi-column network architecture, wherein multiple losses are calculated such that each column converges at different local minima of loss. Consequently, each column extracts different features from the same dataset. Thus, this proposed network architecture is called multi-loss convolutional neural network. This study compares the differences between the proposed multi-loss architecture and existing multi-column architectures. Experiments were conducted using the CIFAR-10 and CIFAR-100 datasets. The results demonstrated that the network architecture proposed in the present study exhibits better performance than those in previous studies; furthermore, the proposed network architecture can be applied to any existing structure, such as ResNet and ResNeXt. Additionally, as with all the multi-column structures, finding a way to aggregate each column remains a problem. Here, a novel method is presented to conduct aggregation in the feature map. Thus, unlike previous studies, the proposed multi-loss architecture can use convolutional operations to obtain higher level features from the combined features. Finally, the correlation between the number of columns and accuracy was investigated. The result showed that the accuracy increased as the number of columns increased until achieving the optimal number.

Keywords convolutional neural network; image classification; multi-column; multi-loss; feature map aggregation

Jian Piao
Doshisha University, Japan, e-mail: ctmd0011@mail4.doshisha.ac.jp

Mingzhe Jin
Doshisha University, Japan, e-mail: mjin@mail.doshisha.ac.jp

TV channels and predictive models: an analysis on social media

Paolo Mariani, Andrea Marletta, Mauro Mussini, and Mariangela Zenga

Abstract In recent years, the collection of data from social network has steeply increased due to the diffusion of internet and portable electronic devices. Data from social network may represent a useful information source to investigate the user opinions on web page contents. Social network users can declare their preferences just by clicking "Like" on a web page. This paper focuses on user's liking for the contents of social network pages by collecting information on the "Likes" given by users to social network pages with a similar content. When considering a set of social network pages with similar contents, the "Likes" given by a user to these pages can be expressed by a binary vector having elements equal to 1 in correspondence of the pages liked by the user and to 0 elsewhere. However, the absence of "Like" on a page can be seen as either a negative opinion or a neutral opinion (lack of knowledge or interest). Thus, the 0 values in the binary vector can be interpreted as a sort of item non-responses, with a single web page being an item. A common approach to missing data treatment consists of substituting them with plausible values. We use a similar approach to deal with missing "Likes" on social network pages to establish whether the user's opinion on these pages is negative or neutral. More specifically, we resort to a missing data imputation procedure to distinguish between the case of a missing "Like" implying a negative opinion ("Dislike") on the page and the case of a missing "Like" meaning a lack of knowledge or interest ("Nothing").

Such a procedure has been applied on Facebook data concerning the official pages of 12 Italian television channels divided into three categories: news, culture and entertainment. The substitution of missing values has been implemented using a threshold based on the number of "Likes" expressed. In particular, the missing value has been translated in a "Dislike" only when users expressed a percentage of "Likes" higher than the selected threshold. Alternatively, if the percentage of "Likes" was lower than the threshold, the missing "Like" was assigned as "Nothing".

Keywords social network; data predictive models; missing values

Paolo Mariani
Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, e-mail: paolo.mariani@unimib.it

Andrea Marletta
Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, e-mail: andrea.marletta@unimib.it

Mauro Mussini
Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, e-mail: mauro.mussini1@unimib.it

Mariangela Zenga
Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, e-mail: mariangela.zenga@unimib.it

Data mining techniques in autobiographical studies. Is there a chance?

Franca Crippa, Angela Tagini, and Fulvia Mecatti

Abstract Unveiling association rules, by means of big when not huge datasets, is at the foundation of data mining techniques. This purpose is particularly relevant in behavioural research, where patterns are intertwined and thus not easily brought to the surface, unless they are trivial. Applications in this direction have tackled behavioural issues in the vast research area of social relations, as in the case of the use of mobile. Our proposal consists in the extension, from the interindividual dimension, to the intraindividual one. Our explorative analysis takes into account episodic memory and imagination, the former in the form of autobiographical retrieval, the latter of simple projection in the future.

We aim at finding frequent itemsets, a base crucial step in data mining. The insight both into the selectiveness of autobiographical memory and into the blind future projection is very demanding and it may not be easily accomplished, resulting possibly in missing data, with open questions on the applicability of resampling techniques.

Keywords data mining; behavioural studies; itemsets

Franca Crippa
Università di Milano-Bicocca, Italy, e-mail: franca.crippa@unimib.it

Angela Tagini
Università di Milano-Bicocca, Italy, e-mail: angela.tagini@unimib.it

Fulvia Mecatti
Università di Milano-Bicocca, Italy, e-mail: fulvia.mecatti@unimib.it

Requirements and competencies for labour market using conjoint analysis

Paolo Mariani, Andrea Marletta, Mauro Mussini, and Mariangela Zenga

Abstract The substantial economic changes affecting the labour market during last years, led to a significant alteration of the requirements requested to the workers. Moreover the success of the firms is not more determined by material resources or huge capitals, but the real key point is represented today by human capital. Therefore, when the guarantee and the certainty of a post is not so indisputable, the worker needs to show the "cross skills", that is personal features of the individual responding to different requests of the labour market. These skills are generally divided into hard and soft, but here a new classification is achieved dividing them into emotional, social and cognitive skills.

These developments led to research being carried out on the requirements of companies that tied themselves in the matching phase to various professional roles, investigating the knowledge, skills, attitudes, and more generally skillsets, by using as evidence the actions meeting supply and demand. Information regarding goodwill, albeit with a managerial and administrative slant, provides a source of knowledge structured on the basis of the criteria that companies adopt in their choices of workers who apply for job positions in their companies. The aim of this work is to measure and evaluate economically the skills that using an a-posteriori analysis of the hired candidates. Considering this context, an application about the classification of the cross skills has been proposed using a Conjoint Analysis in combination with an index of monetary revaluation.

Keywords conjoint analysis; labour market; skills

Paolo Mariani
University of Milano-Bicocca, Italy, paolo.mariani@unimib.it

Andrea Marletta
University of Milano-Bicocca, Italy, andrea.marletta@unimib.it

Mauro Mussini
University of Milano-Bicocca, Italy, mauro.mussini1@unimib.it

Mariangela Zenga
University of Milano-Bicocca, Italy, mariangela.zenga@unimib.it

A data mining framework for Gender gap on academic progress

Adele Marshall, Aglaia Kalamatianou, and Mariangela Zenga

Abstract Gender is considered to have a fundamental influence on research on education. Access and enrolment to higher education and the completion/graduation have their own corresponding importance in higher education research involving gender. The first opens the way for enjoying a public good, but it may not be enough if graduation is not ultimately reached. The research has developed student outcomes, where based on success and student performance or students' efficacy, effectiveness and efficiency. Even though there is no consensus regarding the definition and measurement, those commonly used fit into two categories: degree completion and time-to-degree, more generally length of studies.

The focus of this work is on students' length of studies defined as the time duration between date of first enrolment to a university institution and up to the occurrence of graduation.

Data consists of two individual level data sets derived from social sciences oriented departments in a University in Italy and a University in Greece. We use statistical approach of Survival Analysis and regression tree (in particular MRT, multivariate regression trees) taken from data mining frameworks. Moreover we apply an inequality index, between gender to prove differences in gender for academic students' progression.

Keywords time to graduation; university students; gender parity index

Adele H. Marshall
Queen's University of Belfast, UK, email: a.h.marshall@qub.ac.uk

Aglaia Kalamatianou
Panteion University, Greece, email: aglaiakalamatianou@gmail.com

Mariangela Zenga
University of Milano-Bicocca, Italy, email: mariangela.zenga@unimib.it

Classification through graphical models: evidences from the EU-SILC data

Manuela Cazzaro, Federica Nicolussi, and Agnese Maria Di Brisco

Abstract The European Union Statistics on Income and Living Conditions (EU-SILC) survey aims to gather multidimensional data and to monitor the poverty level, social inclusion, and living conditions of European countries. For the purpose of our study, we focus on evaluating the level of perceived health, which is quantified into five levels from "very good" to "very bad". The sample consists of 345401 adult subjects from 31 European countries. Possible factors that determine the level of perceived health include personal information (age, gender, and level of education), economic status (status in employment, income bracket), and use of free time (capacity to afford paying for one week annual holiday, regular participation in a leisure activity, and access to a small amount of money each week for personal use). These additional variables are dichotomous or categorical ones and all together they constitute a contingency table. To evaluate the relationship between the level of perceived health and the other variables we take advantage of graphical models that offer a rigorous statistical instrument of analysis. Graphical models provide a dependence structure for log-linear models for contingency tables. The (possibly complex) system of dependencies can be easily represented through a graphical representation, where each vertex of the graph corresponds to one variable and an arc between a couple of vertices indicates a dependence. Even more complex structures of dependencies can be well-represented by considering in the same graph both directed and undirected arcs corresponding to symmetric and asymmetric dependencies. In this study we take advantage of the Stratified Chain Regression Graphical models which embody a system of multivariate logistic regressions. In the learning procedure of the best fitting model, we propose to take advantage of a Bayesian learning algorithm, understood as the posterior distribution over graphical models, to select the set of dependencies that best fit the data. The algorithm requires the evaluation of the marginal likelihood, which can be approximated through a maximum likelihood estimation of the Bayesian information criterion score, and of the prior probability of a graph. With respect to the latter issue, we investigate several prior's choice and we favor weakly-informative priors that penalize dense graphs. Finally, we perform a classification algorithm based on classification trees to identify clusters.

Keywords stratified chain regression models; bayesian model learning; perceived health; log-linear model

Manuela Cazzaro
Università degli Studi di Milano-Bicocca, Italy, e-mail: manuela.cazzaro@unimib.it

Federica Nicolussi
Università degli Studi di Milano, Italy, e-mail: federica.nicolussi@unimi.it

Agnese Maria Di Brisco
Università degli Studi di Milano-Bicocca, Italy, e-mail: agnese.dibrisco@unimib.it

Cross-disciplinary higher education of data science – beyond the computer science student

Evangelos Pournaras

Abstract The majority of economic sectors are transformed by the abundance of data. Smart grids, smart cities, smart health, Industry 4.0 impose to domain experts requirements for data science skills in order to respond to their duties and the challenges of the digital society. Business training or replacing domain experts with computer scientists can be costly, limiting for the diversity in business sectors and can lead to sacrifice of invaluable domain knowledge. This paper illustrates experience and lessons learnt from the design and teaching of a novel cross-disciplinary data science course at a postgraduate level in a top-class university. The course design is approached from the perspectives of the constructivism and transformative learning theory. Students are introduced to a guideline for a group research project they need to deliver, which is used as a pedagogical artifact for students to unfold their data science skills as well as reflect within their team their domain and prior knowledge. In contrast to other related courses, the course content illustrated is designed to be self-contained for students of different discipline. Without assuming certain prior programming skills, students from different discipline are qualified to practice data science with open-source tools at all stages: data manipulation, inter- active graphical analysis, plotting, machine learning and big data analytics. Quantitative and qualitative evaluation with interviews outlines invaluable lessons learnt.

Keywords education; data science; cross-discipline; big data; research methodology; learning; constructivism theory; transformative theory

Evangelos Pournaras
ETH Zurich, Switzerland, e-mail: epournaras@ethz.ch

The implications of network science in economic analysis

Éva Kuruczleki

Abstract Using real world data is crucial in economic analysis for students to better understand not only the methodology behind analyses but real world processes. Apart from the use of real data, new analytical methods, such as graph theory and network analysis made their way to economics and statistical education. Educators are facing new challenges on how to introduce new methodological tools in both economic and statistical education. In this study the experiences about the implementation of network analysis and other new methods as part of statistical courses in economics education is introduced, with special focus on analyses concerning European integration processes.

Keywords network analysis; real data; European integration

Éva Kuruczleki
University of Szeged, Hungary, email: kuruczleki.eva@eco.u-szeged.hu

Conception of measures of central tendency of primary school teachers

Evanthis Chatzivasileiou

Abstract In a democracy the need to develop citizens' "sense of data" so that they can convert raw information into organized information sets and make decisions is an objective of minors' education.

The basic concepts of statistics which taught to primary school students are the measures of central tendency (mean, median, mode). According to surveys, students are easily learning the average algorithm, but they do not get to gain the structural and functional understanding of the middle of a distribution.

From research, we found that primary school students have misconceptions in the average as a representative value, fare share, and point of balance of a data distribution. Also they know the algorithm of calculating the mean but not being able to interpret it.

As we try to find out the cause of students' misconceptions, we explored the conception of measures of central tendency of primary school teachers and students of university pedagogical department that prepare future teachers.

The survey revealed interesting observations and findings about teachers' misconceptions about the concepts of measures of central tendency, which could be some of the causes of student misconceptions.

With this announcement we present some of the results of a quantitative survey conducted using questionnaire to 325 primary school teachers and 234 students of the pedagogical department of the Aristotle University of Thessaloniki.

From the results of the survey it appears that the teachers in a significant percentage have misunderstandings corresponding to the students, which may be related to the students' misconceptions after the formal education.

Keywords measures of central tendency

Evanthis Chatzivasileiou
Aristotle University of Thessaloniki, Greece, e-mail: evanthis.chatz@gmail.com

Quality of life profiles of colon cancer survivors: A three-step latent class analysis

Felix J. Clouth, Gijs Geleijnse, Lonneke V. van de Poll-Franse, Steffen Pauws, and Jeroen K. Vermunt

Abstract The aim of this study was to understand better the impact of colon cancer and its treatment on long-term survivor´s lives by determining subgroups (latent classes) among these survivors with different quality of life (QoL) outcomes.

For this study data from the PROFILES 2010 colorectal cancer QoL survey was used. PROFILES is a population-based cohort of cancer survivors aiming to assess the physical and psychosocial impact of cancer and its treatment. Included are patients diagnosed between 2000 and 2009 as registered in the Netherlands Cancer Registry (NCR). The NCR is a nationwide, population-based registry of all newly diagnosed cancer patients in the Netherlands. For this study, survivors with non-metastatic (stage I – III) colon cancer (n = 1510) were selected. Data on patient and tumor characteristics, received treatment, and follow-up survival status of the selected patients were linked from the NCR. Latent class analysis (LCA) was used to identify subgroups with statistically distinct and clinically meaningful QoL patterns. First, based on fifteen QoL indicators as assessed with the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire C30 a standard LCA estimated the classes. Second, class-membership was determined based on the posterior class-membership probabilities using modal assignment and, third, the effect of covariates on class assignment was assessed using multinomial logistic regression. Follow-up survival across classes was compared using Log-Rank tests and Cox regression analyses.

Model fit assessed by the Bayes Information Criterion (BIC) indicates five classes: (a) good overall QoL (41.2%), (b) good physical and medium cognitive functioning (26.8%), (c) medium physical and good cognitive functioning (15.4%), (d) medium overall QoL (9.3%), and (e) poor overall QoL (7.3%). Significant covariates were cancer stage ($p < .001$), gender ($p < .001$), number of comorbidities ($p < .001$), and the interaction of number of comorbidities and received chemotherapy ($p < .001$). 5-year follow-up survival probabilities ranged from .898 in class 1 to .566 in class 5 ($p < 0.001$).

A five-class solution shows differences in responses on the functioning scales for patients with good to medium overall QoL. Apart from this finding, response patterns within classes are surprisingly uniform across the fifteen EORTC dimensions. Our results show that LCA can be a useful tool for better understanding the impact of cancer and its treatment. Our future research will be about how to translate current findings into personalizing care of long-term colon cancer survivors.

Keywords quality of life; colon cancer; three-step latent class analysis; personalized care; EORTC; PROFILES registry; Netherlands cancer registry

Felix J. Clouth
Tilburg University, Comprehensive Cancer Organisation, The Netherlands, e-mail: f.j.clouth@uvt.nl

Gijs Geleijnse
Netherlands Comprehensive Cancer Organisation, e-mail: g.geleijnse@iknl.nl

Lonneke V. van de Poll-Franse
Netherlands Comprehensive Cancer Organisation; The Netherlands Cancer Institute; Tilburg University, l.vandepoll@iknl.nl

Steffen Pauws
Tilburg University, The Netherlands, e-mail: s.c.pauws@uvt.nl

Jeroen K. Vermunt
Tilburg University, The Netherlands, e-mail: j.k.vermunt@uvt.nl

Classifying functional groups of microorganisms with varying prevalence level using 16S rRNA

Rafal Kulakowski, Etienne Low-Decarie, and Berthold Lausen

Abstract The ever-improving sequencing technologies continue to revolutionize our understanding of the importance of microbial communities for human health and the Earth's ecosystem. The 16S ribosomal RNA is a marker gene, commonly sequenced and applied to estimate a phylogenetic make-up of a community. In this study, we investigate the extent to which, 16S rRNA can be used to estimate the functional make-up of a population. The prevailing methods include the use of homological information, however, here we approach the problem as a set of binary classification tasks, where the classifiers predict whether a given individual in a population has a certain functional capability or not. A numerical feature space is derived from the original, sequential data using the Natural Language Processing methods. The imbalanced class distribution hinders the performance of some resulting classifiers, built following basic sampling techniques, especially when used to predict rare functional groups. As part of this investigation, we test a range adjusted sampling techniques for highly imbalanced classes. The results suggest that the misclassification error rate and the precision can be significantly improved for most functional groups, however some of the classes with the lowest prevalence appear to lack adequate representation within the available dataset to train a reliable classifier.

Keywords classification; feature representation; microbial communities; class imbalance

Rafal Kulakowski
Department of Mathematical Sciences, University of Essex, UK, e-mail: rkulaka@essex.ac.uk

Etienne Low-Decarie
School of Biological Sciences, University of Essex, UK, e-mail: etienne.decarie@gmail.com

Berthold Lausen
Department of Mathematical Sciences, University of Essex, UK, e-mail: blausen@essex.ac.uk

Identifying Chronic Obstructive Pulmonary Disease (COPD) phenotypes to predict treatment response

Vasilis Nikolaou, Sebastiano Massaro, Masoud Fakhimi, and Lampros Stergioulas

Abstract Chronic obstructive pulmonary disease (COPD) is a leading cause of death worldwide and a major cause of chronic morbidity and mortality. COPD is a multifaceted disease characterized by persistent respiratory symptoms and airflow limitation which is however preventable and treatable, thereby offering a compelling call to identify novel approaches to optimize patients' responses to treatment. Recent research focusing on patients with acute exacerbations COPD (AECOPD) only, has already provided early evidence that distinctive patient phenotypes are most suited to receive different types of treatment. In this work we substantially extend this evidence by investigating COPD patient phenotypes across the General Practitioners Royal College of General Practitioners (RCGP) Research and Surveillance (RSC) database – one of the oldest and largest sentinel networks in Europe. Specifically, we use both hierarchical and k-means cluster analyses to classify patients according to their physiological and clinical characteristics along with COPD-related comorbidities and blood biomarkers. The resulting clusters allow identifying underlying COPD patient phenotypes and associating them with different treatments. Altogether the framework we put forward here will help identifying targeted and more accurate treatments to improve the management of the disease.

Keywords: COPD phenotypes; hierarchical clustering; k-means clustering; personalized targeted treatment

Vasilis Nikolaou
Surrey Business School, University of Surrey, United Kingdom. E-mail: v.nikolaou@surrey.ac.uk

Sebastiano Massaro
Surrey Business School, University of Surrey, Guildford, United Kingdom, E-mail: s.massaro@ surrey.ac.uk

Masoud Fakhimi
Surrey Business School, University of Surrey, United Kingdom, E-mail: Masoud.fakhimi@surrey.ac.uk

Lampros Stergioulas
Surrey Business School, University of Surrey, United Kingdom, E-mail: l.stergioulas@surrey.ac.uk

The use of gene ontology to improve gene selection process for omics data analysis

Chadia Ed-driouch, Hassan Kafsaoui and Ahmed Moussa

Abstract In Next Generation Sequencing (NGS) Technologies, various researches have been performed to select differentially expressed genes between two or more sample groups and showed beneficial results. Some genes, even if are potentially involved in studied pathology, are unselected because of biological variations (the current state of cell derivation, the chemical signals of other cells, etc.) and/or procedural variations (due to experimentation). This selection mistake could influence the identification of several other genes which might represent new interesting targets. To reveal these genes, we propose a new method to improve the gene selection process which information is coming from gene ontology and gene expression.

The proposed method takes into account gene similarity and semantic similarity, where the first one is related to the NGS experiment and the second one to the gene ontology. This additional a priori information allows to correct the selection process by taking into account functional annotations provided by gene ontology. Experimental results confirm this assumption and show that this new method allow to better understand gene network and explore with precision molecular distinction between health states.

Keywords gene selection; gene ontology; expression similarity

Chadia Ed-driouch
Systems and Data Engineering Team ENSA-Tangier, University Abdelmalek Essaadi, Morocco, e-mail: eddriouch@ensat.ac.ma

Hassan Kafsaoui
Engineering, Environment, Modelisation, and Applications, FS, University Ibn Tofail, Morocco, e-mail: kafssaouih@yahoo.fr

Ahmed Moussa
Systems and Data Engineering Team ENSA-Tangier, University Abdelmalek Essaadi, Morocco, e-mail: amoussa@uae.ac.ma

Properties of individual differences scaling and its interpretation

Niel le Roux, and John Gower

Abstract Indscal models consider symmetric matrices $\boldsymbol{B}_k = \boldsymbol{X}\boldsymbol{W}_k\boldsymbol{X}'$ for $k = 1, \ldots, K$, where $\boldsymbol{X}$ is a compromise matrix termed the group-average and $\boldsymbol{W}_k$ is a diagonal matrix of weights given by the kth individual to the R, specified in advance, columns of $\boldsymbol{X}$; non-negative weights are preferred and usually R < n. We propose a new two-phase alternating least squares (ALS) algorithm, that emphasizes the two main components (group average and weighting parameters) of the Indscal model and helps with the interpretation of the model. Furthermore, it has thrown new light on the properties of the converged solution, that would be satisfied by any algorithm that minimizes the basic Indscal criterion: $Min \sum_{k=1}^{K} \|\boldsymbol{B}_k - \boldsymbol{X}\boldsymbol{W}_k\boldsymbol{X}'\|^2$ where the minimization is over $\boldsymbol{X}$ and the $\boldsymbol{W}_k$ The new algorithm has also proved to be a useful tool in unravelling the algebraic understanding of the role played by parameter constraints and their interpretation in variants of the Indscal model. The proposed analysis focusses on Indscal but the approach may be of more widespread interest, especially in the field of multidimensional data analysis. Some of the main issues are: Simultaneous least-squares estimates of the parameters may be found without imposing constraints. However, group average and individual weighting parameters may not be estimated uniquely, without imposing some subjective constraint that could encourage misleading interpretations. We encourage the use of linear constraints $\boldsymbol{W}_k\mathbf{1} = \mathbf{1}$, $k = 1, \ldots, K$, as it enables a comparison of the weights obtained (i) within group k and (ii) between the same item drawn from two or more groups. However, it is easy to exchange one system of constraints to another in a post- or pre-analysis. The new two-phase ALS algorithm (a) computes for fixed $\boldsymbol{X}: n \times R$ the weights $\boldsymbol{W}_k$ subject to $\boldsymbol{W}_k\mathbf{1} = \mathbf{1}$, and then (b) keeping $\boldsymbol{W}_k$ fixed, it updates $\boldsymbol{X}$. At convergence, the estimates of $\boldsymbol{X}: n \times R$ and the $\boldsymbol{W}_k$ will apply to all algorithms that minimize the Indscal criterion. Furthermore, we show that only at convergence an analysis-of-variance property holds on the demarcation region between over- and under-fitting. When the analysis-of-variance is valid, its validity extends over the whole matrix domain, over trace operations, and to individual matrix elements. The optimization process is unusual in that optima and local optima occur on the edges of what seem to be closely related to Heywood cases in Factor analysis.

Keywords indscal; alternating least squares; group average

Niel le Roux
Stellenbosch University, South Africa, e-mail: njlr@sun.ac.za

John Gower
The Open University, U.K., e-mail: john.gower@open.ac.uk

Local and global relevance of features in multi-label classification

Trudie Sandrock

Abstract Multi-label classification problems arise in scenarios where every data instance can be associated simultaneously with more than one label. Feature selection in a multi-label context is more challenging than in the single label case, since additional complexity is introduced by the fact that features which may discriminate well between values of one of the labels will not necessarily do the same for other labels. In this regard the concepts of local and global relevance of features will be introduced. A multi-label feature selection procedure should take cognisance of the possibility that some features may not be globally relevant, but could be locally relevant for one or more labels. I propose a new multi-label feature selection method, based on a binary relevance problem transformation. Empirical results obtained from applying the proposed technique as well as existing techniques to benchmark datasets will be reported. These results show that performing feature selection at a local level leads to improvements in performance and also simultaneously provides additional insight into the data.

Keywords feature selection; multi-label classification; local and global relevance

Trudie Sandrock
Stellenbosch University, South Africa, e-mail: trudies@sun.ac.za

A multivariate ROC based classifier

Martin Kidd

Abstract Receiver Operating Curve (ROC) analysis is well known, especially in biostatistics for determining optimal cut-off values of continuous variables for predicting "positive" vs "negative" outcomes of a binary dependent variable. This single variable ROC analysis is extended to a multivariate ROC classifier consisting of p predictor variables each with corresponding "optimal" cut-off value. A further ROC analysis is conducted on the number of positive outcomes (or signs) from the p predictor variables to determine an optimal threshold k. A positive classification for a single case is made if there are k or more signs across the p predictor variables.

The fitting of such a multivariate ROC model is a simple two step procedure of firstly obtaining optimal cut-off values from a training dataset for each predictor variable in the sense of optimising sensitivity and specificity simultaneously. An added feature of this optimisation is that restrictions could be placed on the sensitivity/specificity levels, i.e. sensitivity greater than a clinically specified level (eg>0.9). In the second step, the number of positive signs are calculated for each training case based on the cut-off values determined in the first step. A further ROC analysis is conducted to determine k, the optimal number of positive signs that need to be present for a positive classification.

Simulated data has indicated that the performance of this classifier could be negatively affected by redundant predictor variables that make no contribution to the classification. Variable selection is therefore important to remove as many redundant predictors as possible. The non-dominated sorting genetic algorithm (NSGA) is applied in this regard to simultaneously search for optimal predictors together with their optimal cut-off values. Linear programming can also be used to jointly determine relevant predictors and corresponding cut-off values.

The multivariate ROC classifier will be demonstrated by applying it to genetic marker data for predicting TB, and producing easy-to-understand prediction models that are appealing to clinicians. Another characteristic of the ROC classifier which makes it suitable for genetic marker data is that it is not affected by outliers in the data.

The ROC classifier will be compared with more complex models like discriminant analysis to show comparable performance.

Keywords Receiver Operating Curve; ROC, Sensitivity; Specificity; Non-dominated Sorting Genetic Algorithm; NSGA; Linear Programming

Martin Kidd
University of Stellenbosch, South Africa, email: mkidd@sun.ac.za

Functional linear discriminant analysis for several functions and more than two groups

Sugnet Lubbe

Abstract Canonical variate analysis in a multivariate data analysis setting aims to find a linear combination of p variables which, after transformation to the canonical space, are optimally separated among k group. First focussing on the two-group case, a single canonical variate is defined maximising the between group relative to within group variance ratio. Many functional data analysis methods are based on multivariate data methods where instead of dealing with p variables, continuous functions are a generalisation with $p \to \infty$. Functional linear discriminant analysis as such a generalisation have been discussed by several authors in the literature. Another point of view is to have p continuous functions and to search for a linear combination of the p functions, such that the resulting functions are optimally separated in the canonical function space.

In this paper a new suggestion for the latter problem is proposed. Two possible solutions are evaluated finding a single set of p coefficients to perform the canonical transformation, or finding time-varying coefficients, i.e. functions of coefficients to transform each time point to a canonical functional space. Furthermore, both these methods can be generalised to discriminant analysis for $k>2$ groups. An optimal two- or three-dimensional visualisation of the canonical functional space is constructed and illustrated with an example.

Keywords linear discriminant analysis; functional data analysis; classification

Sugnet Lubbe
Stellenbosch University, So3uth Africa, e-mail: slubbe@sun.ac.za

Variable selection in linear regression models with non-gaussian errors: a bayesian solution

Giuliano Galimberti, Saverio Ranciati, and Gabriele Soffritti

Abstract Most of the inferential procedures associated with linear regression models rely on the assumption that (i) the error terms follow a Gaussian distribution; (ii) the model contains all the relevant regressors that affects the outcome. These assumptions can be violated in many practical applications. For example, the error distribution might be characterised by heavy tails, skewness and/or multimodality. Furthermore, the set of relevant regressors may not be known in advance, but its identification can be part of the investigation process.

To simultaneously address departures from Gaussian distribution and variable selection, a Bayesian solution is proposed. In particular, a class of linear regression models with errors distributed according to a mixture of Gaussian distributions is defined, with the choice of relevant regressors embedded in the model specification. Two layers of unobservable latent variables are introduced, thus leading to a hierarchical formulation of such models. A weakly informative modified *g*-prior for the regression coefficients is elicited, that is a conjugate prior for the likelihood of mixture model's component. This prior also induces a form of penalization, thus overcoming potential problems of overfitting. Conjugate prior distributions for the remaining parameters are also proposed, resulting in closed form conditional posterior distributions. Exploiting this formulation, parameter estimation and variable selection are performed simultaneously, by sampling from the posterior distribution associated with the model. In particular, a Monte Carlo Markov Chain implementation of the sampling procedure is derived, consisting of Gibbs samplers steps based on full conditionals for the model parameters. Since the number of components is held fixed in this MCMC algorithm, the Deviance Information criterion is proposed as a tool to select the optimal number of components.

The proposed methodology is compared with other variable selection techniques through an extensive simulation study, in order to evaluate the impact on their performances due to different sources of deviation from normality. The effects of other quantities such as the sample size and the number of relevant/candidate covariates are also assessed. In particular, this simulation study show that the proposed methodology appears to be effective in selecting the relevant regressors when the distribution of the error terms is characterised by heavy tails, skewness and/or multimodality. Results obtained on a real data set are also described.

Keywords mixture models; modified *g*-prior; MCMC

Giuliano Galimberti
Università di Bologna, Italy, e-mail: giuliano.galimberti@unibo.it

Saverio Ranciati
Università di Bologna, Italy, e-mail: saverio.ranciati2@unibo.it

Gabriele Soffritti
Università di Bologna, Italy, e-mail: gabriele.soffritti@unibo.it

Finite mixtures of matrix-variate regressions with random covariates

Salvatore Daniele Tomarchio, Paul D. McNicholas, and Antonio Punzo

Abstract Finite mixtures of regressions are a well-known model-based clustering technique for dealing regression data. However, they assume assignment independence, i.e., the allocation of data points to the clusters is required to be independent from the covariates distribution. This assumption is generally not true and makes inadequate the models in many real data applications. To overcome this problem, finite mixtures of regressions with random covariates, also known as cluster-weighted models (CWMs) have been proposed in the literature. In this paper, a matrix-variate CWM is introduced. Specifically, both the response conditional means and the covariates distribution are assumed to be matrix normal. The model can be thought as a generalization of a multivariate-multiple CWM, in which the sets of variables are simultaneously observed at different time points, resulting in a three-way data structure. Maximum likelihood parameter estimates are derived using the EM algorithm. Classification assessment and clustering results are analyzed on simulated and real data.

Keywords matrix-variate regression; cluster-weighted model

Salvatore D. Tomarchio
Università degli studi di Catania, Italy, e-mail: daniele.tomarchio@unict.it

Paul D. McNicholas
McMaster University, Canada, e-mail: paulmc@mcmaster.ca

Antonio Punzo
Università degli studi di Catania, Italy, e-mail: antonio.punzo@unict.it

Telescoping mixtures - Learning the number of components and data clusters in Bayesian mixture analysis

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün

Abstract Telescoping mixtures are an extension of sparse finite mixtures by assuming that additional to the unknown number of data clusters also the number of mixture components is unknown and has to be estimated. Telescoping mixtures explicitly distinguish between the number of data clusters K_+ and components K in the mixture distribution, and purposely allow for more components than data clusters. By linking the prior on the number of components to the prior on the mixture weights, it is guaranteed that components remain empty as K increases, making the number of clusters in the data, defined through the partition implied by the allocation variables, random a priori. Telescoping mixtures can be seen as an alternative to infinite mixtures models. We present a simple algorithm for posterior MCMC sampling to jointly sample K, the number of components, and K_+, the number of data clusters. The algorithm is compared to standard transdimensional algorithm such as the reversible jump Markov chain Monte Carlo and the Jain-Neal split-merge sampler.

Keywords finite mixture; unknown number of components; sparse finite mixture

Gertraud Malsiner-Walli
WU Vienna University of Economics and Business, Austria, e-mail: gmalsine@wu.ac.at

Sylvia Frühwirth-Schnatter
WU Vienna University of Economics and Business, Austria, e-mail: sfruehwi@wu.ac.at

Bettina Grün
Johannes Kepler Universität Linz, Austria, e-mail: bettina.gruen@jku.at

Finite mixture modeling and model-based clustering for directed weighted multilayer networks

Volodymyr Melnykov, Shuchismita Sarkar, and Yana Melnykov

Abstract A novel approach relying on the notion of mixture models is proposed for modeling and clustering directed weighted networks. The developed methodology can be used in a variety of settings including multilayer networks. Computational issues associated with the developed procedure are effectively addressed by the use of MCMC techniques. The utility of the methodology is illustrated on a set of experiments as well as applications to real-life data containing export trade amounts for European countries.

Keywords model-based clustering; directed network; weighted network; multilayer network; MCMC

Volodymyr Melnykov
University of Alabama, USA, e-mail: vmelnykov@cba.ua.edu

Shuchismita Sarkar
Bowling Green State University, USA, e-mail: shuchismita.sarkar@gmail.com

Yana Melnykov
University of Alabama, USA, e-mail: ymelnykov@cba.ua.edu

Intertemporal exploratory analysis of Greek households in relation to information and communications technology (ICT) from official statistics

Stratos Moschidis, and Athanasios C. Thanopoulos

Abstract The continued development of information and communication technologies has shaped their use by households respectively. The present paper aims to map the effects of these changes through an exploratory analysis of the phenomena trends. For the implementation of this study data from Official Statistics were used, as they were recorded through sample surveys of the Greek Statistical Authority (ELSTAT) during the period from 2008 to 2018. This work is of multiple interest. This is because, on the one hand, the phenomenon itself is an area of general scientific interest, but also because the period of data collection includes the time horizon of the beginning of the economic crisis in Greece. The results of the study constitutes the trigger of the debate on how austerity has influenced different practices on this particular issue.

Keywords official statistics; multiple correspondence analysis; multivariate statistics

Stratos Moschidis
Hellenic Statistical Authority, Greece, e-mail: smos@statistics.gr

Athanasios C. Thanopoulos
Hellenic Statistical Authority, Greece, e-mail: a.thanopoulos@statistics.gr

Hierarchical clustering for anonymization of economic survey data

Kiyomi Shirakawa, and Takayuki Ito

Abstract In data anonymization, statisticians in government authorities usually apply top or bottom coding, which ensures the confidentiality of information in the secondary use of microdata. Specifically, a top-code (bottom-code) gives an upper (lower) bound for a variable. Any observation beyond the threshold should be encompassed by a categorical indicator or intentionally concealed in the microdata file. Obviously, the manipulated data lose the information on the tail of distribution, thus causing handling difficulties from a data-user perspective. In order to overcome this problem, in 2018, we developed an R program which allows for constructing practically appropriate threshold values for top and bottom coding. In particular, this program searches for the borders of groups in a given dataset with the decision tree method, and gives possible options of the lower bound. The program then identifies the suitable bound with the smallest AIC of the Chow test.

This study further extends our top/bottom coding method using the multi-level models. In fact, in our last study, we did not consider the case that end users can identify specific companies with large sales, which should be prevented by the data administrator for the purpose of confidentiality. Also, we just simply applied regression models in a single year. Therefore, in this study we illustrate our top-coding method with a longitudinal corporate data. In doing so, we are able to avoid the identification of specific conglomerates. Specifically, we use EDINET (Electronic Disclosure for Investors' NETwork), which publishes corporate data, and created a regression model for sales at two levels per industry classification and corporate. Furthermore, we add time point data and created a multi-level model of 3-level hierarchy, i.e., industry-level, corporate-level, time-level.

Keywords secondary use; top coding; multi-level model; official statistics; longitudinal data

Kiyomi Shirakawa
Hitotsubashi University, Japan, e-mail: kshirakawa@ier.hit-u.ac.jp

Takayuki Ito
Hitotsubashi University, Japan, e-mail: tito4@ier.hit-u.ac.jp

Improvement of training data based on pattern of reliability scores for overlapping classification

Yukako Toko, Mika Sato-Ilic, and Shinya Iijima

Abstract We developed the supervised multiclass classifier for automated coding in our previous studies. The developed classifier assigns classification codes based on reliability scores. The previously defined reliability score considers both the uncertainty from data (probability measure) and the uncertainty from latent classification structure in data (fuzzy measure) in order to consider the unrealistic restriction of an ambiguous text description being classified to a single class, which gives our method a better accuracy of the result. Using this reliability score, multiple classification codes can be assigned corresponding to a text description. From our experiences, this classifier could obtain higher accuracy when compared with our prior studies of classifier without reliability scores.

In this case, we used full data included both clear and unclear data in which clear data means experts can easily assign a text description to a single code. Otherwise, unclear data means experts have difficulty in assigning a text description to a single code. For our current study, target data is only unclear data. The task of getting better accuracy for the result is much harder. Therefore, the purpose of this study is the improvement of training data based on the pattern of a reliability score to improve the classification accuracy.

First, we classify the data of reliability scores to obtain patterns of reliability scores of text descriptions. Next, we detect text descriptions belonged to clusters whose reliability scores are relatively small. That is, we capture text descriptions which do not belong to any classification codes. In other words, we assume that such text descriptions are rarely occurring text descriptions, and try to add the information of the rarely occurred text descriptions to the original training data set which indicates the supervisor of this classifier to improve the training data set. We implement a classifier that includes the improved training data set. In our presentation, we show a better performance of the proposed classifier involved the improved training data set.

Keywords Coding; improvement training data set; pattern of reliability scores

Yukako Toko
National Statistics Center, Japan, e-mail: ytoko@nstac.go.jp

Mika Sato-Ilic
University of Tsukuba, National Statistics Center, Japan e-mail: mika@risk.tsukuba.ac.jp

Shinya Iijima
National Statistics Center, Japan, e-mail: siijima@nstac.go.jp

The epistemology of nondistributive profiles

Patrick Allo

Abstract Automated profiling is a knowledge-discovery method that is used to classify persons and inform decisions about these persons on the basis of insights and patterns that can be discovered within data-sets. So-called nondistributive profiles are profiles that rely on non-universal generalisations for the classification of persons. Current evaluations of the ethical and epistemological risks that are associated with automated profiling-practices often rely on the associated distinction between distributive and nondistributive profiles to explain those risks. Specifically, the diagnosis that nondistributive profiles may coincidentally situate an individual in the wrong category is often perceived as the central shortcoming of such profiles. According to this diagnosis, most risks can be retraced to the inevitability of false positives and false negatives. This article develops an alternative analysis of nondistributive profiles in which this fallibility of nondistributive profiles is no longer the central concern. Instead, it focuses on how profiling creates various asymmetries between an individual data-subject and a profiler. The emergence of informational, interest, and perspectival asymmetries between data-subject and profiler explains how nondistributive profiles weaken the epistemic position of a profiled individual. This alternative analysis provides a more balanced assessment of the epistemic risks associated with nondistributive profiles.

Keywords profiling; nondistributive profiles; rationality; data-ethics; method of abstraction

Patrick Allo
Vrije Universiteit Brussel, Belgium, e-mail: patrick.allo@vub.be

Prediction without estimation: a case study in computer vision

Jérémy Grosman

Abstract The paper describes some of the minute actions through which engineers working in computer vision build systems capable of discriminating between human individuals or activities and ensure the stabilities of their capacities across a wide range of situations. The brief empirical account of these practices precedes a more conceptual discussion about classification practices and, more specifically, about classification practices that do not rely on more traditional forms of statistical estimation.

The paper consequently focuses on the three main actions and arguments engineers usually invoke for supporting their systems' performances: (i) the careful organization of data collections (i.e. engineers in computer vision often have to generate data about the individuals or activities they wish to discriminate), (ii) the extraction of sound visual features prior processing sequences of images (i.e. engineers in computer vision, at least prior the advent of 'deep learning', heavily relied on such visual descriptors) and (iii) the robustness of the metrics used for training, validating and evaluating their models (i.e. engineers in computer vision jointly appeal to the reliability of 'cross-validation' methods and the system's performances across various metrics).

The empirical material essentially consists – besides the usual technical literature – of interviews and observations conducted during FP7 and H2020 projects, respectively titled 'Privacy Preserving Perimeter Protection Project' (P5: 2013–2016) and 'Pervasive and User Focused Biometrics Border Project' (PROTECT: 2016–2019). The overall research has more been supported by the FNRS through a research project titled 'Algorithmic Governmentality' (2013 – 2017).

Keywords history and philosophy of science; science and technology studies; machine learning; computer vision

Jérémy Grosman
Université de Namur, Belgium, e-mail: jeremy.grosman@unamur.be

Reconceptualizing null hypothesis testing

Jan Sprenger

Abstract There is a widespread feeling that scientific method is in a state of crisis, at witnessed by the replication crisis haunting various scientific disciplines. Among the manifold causes of this crisis, I would like to focus (yet again) on a specific aspect of statistical methodology: the use of null hypothesis significance testing (NHST). The various reform proposals that address the shortcomings of NHST fall into three categories: (1) they discard the idea of hypothesis testing altogether (e.g., by shifting to estimation or descriptive statistics), (2) they replace (asymmetric) NHST by symmetric Bayesian hypothesis testing, or (3) they leave the basic rationale of NHST intact and try to cure the most egregious problems in an ad hoc manner (e.g., by lowering the statistical significance threshold to $p < 0.005$).

I argue that none of these proposals is satisfactory. Instead, we need to rethink the epistemic goals of hypothesis testing in science and to reconceptualize our statistical procedures accordingly. I will show that the thoughts of the philosopher Karl R. Popper are useful for developing a foundationally sound and practically applicable method of hypothesis testing, based on the concept of degree of corroboration of the null hypothesis. Then I show how this method improves upon the shortcomings of the above proposals while saving important intuitions from both Bayesian and frequentist reasoning.

Keywords statistics; hypothesis testing; objectivity; Bayesian reasoning; corroboration

Jan Sprenger
University of Turin, Italy, e-mail: jan.sprenger@unito.it

Progress of statistics and data science education in Japanese universities

Akimichi Takemura

Abstract In April of 2017 Shiga University launched an undergraduate program in the new faculty of data science, which is the first one in Japan. Then it launched a master program in data science in April of 2019. This faculty emphasizes statistics and it can be regarded also as the first faculty of statistics. The inauguration of the faculty marks a new era of statistics and data science education in Japanese universities, in view of the fact that there were virtually no faculty of statistics in Japanese universities before Shiga University. In April of 2018 Yokohama City University followed Shiga University with its new school of data science. In April 2019 two other universities followed with similar faculties. We now see rapid progress of statistics and data science education in Japan. We discuss these developments and prospects of statistics and data science in Japan.

Keywords big data; domain knowledge

Akimichi Takemura
Shiga University, Japan, e-mail: a-takemura@biwako.shiga-u.ac.jp

Before Teaching Data Science, Let's First Understand How People Do It

Rebecca Nugent

Abstract In the last few years, the number of (primarily graduate) programs in Data Science has grown to the hundreds. Most of these programs were built on a foundation of already existing courses in several computing-oriented departments; less effort, understandably under the constraint of development efficiency, has been spent on understanding the integration of all of the necessary skills or how people from diverse backgrounds and disciplines approach or think about data science. The Department of Statistics & Data Science at Carnegie Mellon is inside the Dietrich College of Humanities and Social Sciences. In addition, our undergraduate program teaches about a third of the campus population every semester (Statistics, Math, Computer Science, Business, etc), so our sequences are taken by hundreds of students with incredibly diverse future degrees ranging from English Rhetoric to Chemistry to Statistics & Machine Learning. We are in an excellent position to characterize how students with very diverse backgrounds approach or even think about Data Science. We have designed and built ISLE (Interactive Statistics Learning Environment), an interactive platform that removes the computing cognitive load and lets students explore Statistics & Data Science concepts in both structured and unstructured ways. The platform also supports student-driven inquiry and case studies. We track every click, word used, and decision made (e.g., which graphs are designed/explored before settling on a final histogram) throughout the entire data analysis pipeline from loading the data to the final written report. Models of the students' online behavior and decisions also include performance metrics as well as what areas they're choosing to study. The platform is flexible enough to allow adaptation, providing different modes of data analysis instruction, active learning opportunities, and exercises for different subsets of the population. Students are also able to build their own case studies with little restriction or faculty intervention. Teaching Data Science while simultaneously learning how we do it.

Rebecca Nugent
Carnegie Mellon Statistics & Data Science, United States, e-mail: rnugent@stat.cmu.edu

Simultaneous clustering and dimension reduction on multi-block data

Shuai Yuan, and Katrijn Van Deun

Abstract Thanks to the trend of gradual adaption of data-rich research, social and behavioral studies more and more often yield multi-block data, which contains different types of measurements collected from the same sample. Such multi-block data are especially interesting and informative since they may contain *common variation*, the covariation between variables of each and every data block, which suggests complex social mechanisms where several factors act jointly. In real data of interest to applied researchers, often heterogeneity in such *common variation* is present but there is no prior knowledge yet on either the structural common variation or the subgroups differences therein. Two existing approaches to this problem are a tandem analysis of the multi-block data (e.g. iCluster, moCluster) and simultaneous clustering and dimension reduction of the concatenated data (e.g. Reduced *K*-means, Factorial *K*-means or Factor Discriminant *K*-means). However, both approaches are problematic in analyzing multi-block data. On the one hand, the tandem analysis (i.e. sequential dimension reduction and *K*-means clustering) may extract lower dimensions that are not informative at all for the subsequent clustering analysis. On the other hand, the application of simultaneous clustering and dimension reduction methods on the concatenated data does not account for the multi-block structure of the data and the presence of strong sources of block-specific variation that obscures the common variation, which is typically of primary interest.

Therefore, a simultaneous procedure is needed that accounts for the multi-block structure of the data and the co-presence of common and block-specific variation therein, and that achieves clustering and dimension reduction at the same time.

In the current paper, we propose such a method. The method involves two steps: in the first step, regularized SCA is applied to filter out the distinctive variation; in the second step, a modified version of FDKM is applied to achieve simultaneous clustering and dimension reduction. The modified version has the additional benefit of performing variable selection, which might be especially useful to interpretation in the presence of high-dimensional data. Furthermore, the performance of the new approach is compared with several alternatives, including the tandem analysis methods on multi-block data (iCluster and moCluster) and the application of traditional *K*-means, RKM, FKM and FDKM on the concatenated data.

Keywords Integrative clustering; dimension reduction; simultaneous component analysis

Shuai Yuan
Tilburg University, Netherlands, email: s.yuan@uvt.nl

Katrijn Van Deun
Tilburg University, Netherlands, email: K.VanDeun@uvt.nl

Model-based hierarchical parsimonious clustering and dimensionality reduction

Carlo Cavicchia, and Maurizio Vichi, and Giorgia Zaccaria

Abstract The development of new technologies is bringing about a huge amount of information, both from the statistical units and variables point of view. On the one hand, cluster analysis techniques provide methods to pinpoint homogeneous groups of objects; on the other hand, dimensionality reduction techniques are useful to tackle the complexity problem and to identify, where possible, latent concepts underlying the observed variables. In both cases, a hierarchical structure of the data cannot be investigated, whenever it exists, since the relationships between clusters or latent constructs are not highlighted.

Furthermore, the growing dissemination of information induces the so-called *big data*, which frequently need a reduction in terms of objects and variables in order to extract relevant information. In many situations, this reduction of the two data dimensions (objects and variables) is obtained applying a factorial method and, sequentially, a cluster analysis. Therefore, this approach, called tandem analysis, usually entails masking the clustering structure of the data.

Parsimonious methods have been developed to simplify the complete hierarchies, building trees with a reduced number of internal nodes aiming at a better interpretation of the results. This objective is often required in policy-making, where the concern is particularly directed to investigate the best performance units with respect to a latent dimension with a broad economic or social impact.

In this paper, starting from a data matrix **X** of size (n × J), with n objects and J quantitative manifest variables (MVs), we propose a model to obtain a simultaneous hierarchical clustering of the statistical units – aggregated around centroids – and dimensionality reduction of the MVs via components. This model-based approach aims at identifying the highest number of clusters with statistically non-significant differences between objects and the minimum number of unidimensional components, from which building two hierarchies. These latter are defined starting from a *Clustering and Disjoint Principal Component Analysis* solution, with a fixed number of clusters and latent concepts, up to a unique broader group of objects and a measure of synthesis, often referred to as a *composite indicator*, for features. The approach is simultaneous; thus, it allows to obtain components of maximum variance in a reduced space of centroids.

The proposed model is estimated in a least-squares semi-parametric framework and some interesting properties are given. A coordinate descent algorithm is provided; although we face with a partitioning problem belonging to the NP-hard class, it turns out to be efficient in real applications.

Keywords clustering; dimensionality reduction; hierarchical models; k-means; disjoint principal component analysis

Carlo Cavicchia
University of Rome La Sapienza, Italy, e-mail: carlo.cavicchia@uniroma1.it

Maurizio Vichi
University of Rome La Sapienza, Italy, e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome La Sapienza, Italy, e-mail: giorgia.zaccaria@uniroma1.it

Active labeling using model-based classification

Cristina Tortora

Abstract Active labeling refers to the interaction of the user with a learning algorithm to manually label some data points. The goal is to show as few instances as possible to the user while obtaining high accuracy, i.e. correct labels.

This problem is very common when classifying text data. We propose a solution that involves two statistical techniques, first, text analysis and dimension reduction, second, cluster analysis and classification using mixtures of generalized hyperbolic distributions (MGHD). At first the textual data are analyzed using Global Vectors for Word Representation (GloVe) or its extension SpaCy. This first step transforms the text data into numerical on which we perform cluster analysis using MGHD. The user will then manually label few observations per cluster randomly selected around the center of each cluster. This first step insure that all the clusters are correctly centered. Using these labels, the second classification step is performed and the user is asked to manually label points that are in between clusters.

The output is a completely classified dataset with a reduced effort from the user. The technique is compared with the most commonly used techniques in active labeling.

Keywords active labeling; model-based classification; generalized hyperbolic distribution; text data

Cristina Tortora
San Jose State University, CA, USA, e-mail: cristina.tortora@sjsu.edu

Chunk-wise PCA with missings

Alfonso Iodice D'Enza, Angelos Markos, and Francesco Palumbo

Abstract In data analytics applications, it is common to deal with incomplete observations, that is, with data sets with missing values. Many supervised and unsupervised learning methods, however, cannot be applied to incomplete data in a seamless way, and principal component analysis (PCA) makes no exception. Traditional strategies to apply PCA to incomplete data sets include: i) to remove observations that contain at least one missing value, with the obvious drawback of possibly discarding a considerable amount of data; ii) to impute the missing entries before applying PCA. A further strategy is to obtain a PCA solution of incomplete data by skipping the missing entries. In the literature, however, PCA algorithms properly designed to deal with incomplete data have also been proposed. Among extant methods, an iterative algorithm for regularised PCA was shown to outperform alternative approaches in recent comparative reviews. One aspect characterising the majority of PCA methods on missing data is their implementation being based on iterative procedures: this is not desirable when dealing with "tall" incomplete data sets, where "tall" refers to data sets with a large number of observations. The aim of this paper is to provide a chunk-wise implementation of the iterative PCA-based imputation strategy, that is suitable for tall data sets and that is able to impute each upcoming chunk based on the previously analysed data. The proposed procedure is compared to its batch counterpart and to a naive implementation that imputes each data chunk independently. A series of experiments were conducted to investigate the performance of the proposed approach, taking into account different missing data mechanisms.

Keywords missing data; PCA, tall data; unsupervised learning

Alfonso Iodice D'Enza,
Dipartimento di Scienze Politiche Universita degli studi di Napoli Federico II, Italy, e-mail: iodicede@unina.it

Angelos Markos
Department of Primary Education, Democritus University of Thrace, Greece, e-mail: amarkos@eled.duth.gr

Francesco Palumbo
Dipartimento di Scienze Politiche Universita degli studi di Napoli Federico II, Italy, e-mail: fpalumbo@unina.it

Multidimensional data analysis of shopping records towards knowledge-based recommendation techniques

George Stalidis, Pantelis Kaplanoglou, and Kostas Diamantaras

Abstract The digitization of retail sales has already given rise to a wave of intelligent web and mobile applications. In this context, personalized recommendations for products, offers or gifts is a valuable marketing tool, aiming at business goals such as up-selling, customer satisfaction, retention and increased cash flow. Modern recommendation systems aim at learning customers' behavior and providing the best personalized recommendations, far beyond simple techniques such as association rule mining and market segmentation based on statistics. The aim of this work is to extract knowledge from supermarket purchase records, reflecting the ordering profiles of individual customers, as a first step towards a knowledge-based recommendation mechanism. The ultimate goal is to use the results of this work as a component of an advanced intelligent recommendation system for super market chains, which combines machine learning with business rules in a complex knowledge model. The envisaged system will be used to produce personalised promotion notifications on special offers, optimized regarding relevance, novelty, serendipity and diversity. The specific goal within this paper was to investigate the ability of multidimensional data analysis to predict from a series of recorded online orders, the future purchases of individual supermarket customers. Multiple Correspondence Analysis and Hierarchical Clustering methods were used to explore ordering habits, to identify clusters of customers based on their purchasing behavior and to associate individual customers with specific preferences. The methods were applied on the "Instacart Online Grocery Shopping Dataset 2017". Input data included sets of 4 to 100 online orders for each individual customer across 200.000 users, where, apart from their order history, customers were unknown. The available data, which were used to compose feature vectors, were products per order, products' departments, order's day of week and time of day, as well as days passed since previous order. The output was twofold: to predict the latest order of each customer, given his order history and to obtain insights in ordering behavior that may lead to the building of marketing rules. The factor and clustering analysis resulted in representative classes of orders, including both typical and niche ones, and revealed associations with prediction value among customers, order content and order parameters.

Keywords recommendation systems; multidimensional data analysis; intelligent notifications

George Stalidis
International Hellenic University, Greece, e-mail: stalidgi@mkt.teithe.gr

Pantelis Kaplanoglou
International Hellenic University, Greece, e-mail: pikaplanoglou@gmail.com

Kostas Diamantaras
International Hellenic University, Greece, e-mail: kdiamant@it.teithe.gr

Principal Component Analysis to explore social attitudes towards the green infrastructure plan of Drama city

Vassiliki Kazana, Angelos Kazaklis, Dimitrios Raptis, Efthimia Chrisanthidou, Stella Kazakli and Nefeli Zagourgini

Abstract A complex green infrastructure plan has been recently put into operation in the city of Drama located in north eastern part of Greece, which aims at the upgrading of environmental and bioclimatic conditions of the city down town area. Within the project a governance network has been established to promote active social participation and increase the project's social acceptability. This work presents the preliminary results of the first of a series of social surveys carried out within the governance network's function to explore the attitudes of the entrepreneurs of the area, who are directly affected by the project. A total of 117 responses were collected by using a specifically designed questionnaire and Principal Component Analysis was conducted to identify the main factors influencing the entrepreneurs' attitude towards the green infrastructure project. These factors involve the construction of water supply and sewerage network, the construction of underground power supply cables, network lighting, road and pavement reconstruction, tree planting, special bioclimatic constructions, traffic signs, aesthetic improvement, quality of life improvement, number of customers increase and turnover increase after project completion, number of customers decrease, turnover decrease and increase of dust during project implementation. Through clustering three groups of entrepreneurs were identified in terms of their attitude towards the green infrastructure project: a) utilitarians, b) positive to change and c) negative to change. Each group was profiled according to demographic characteristics and awareness preferences.

Keywords principal component analysis; clustering; governance; green infrastructure; social attitudes

Vassiliki Kazana
Eastern Macedonia & Thrace Institute of Technology, Greece, e-mail: vkazana@teiemt.gr

Angelos Kazaklis
Olympos Non Profit Integrated Centre for Environmental Management, Greece, e-mail: akaz98@otenet.gr

Dimitrios Raptis
Eastern Macedonia & Thrace Institute of Technology, Greece, e-mail: d_rapt@yahoo.gr

Efthimia Chrisanthidou
Eastern Macedonia & Thrace Institute of Technology, Greece, e-mail: efchrisanthidoy@gmail.com

Stella Kazakli
Eastern Macedonia & Thrace Institute of Technology, Greece, e-mail: stkaz98@gmail.com

Nefeli Zagourgini
Eastern Macedonia & Thrace Institute of Technology, Greece, e-mail: nefeliza@gmail.com

MCA's visualization techniques: an application in social data

Vasileios Ismyrlis, Efstratios Moschidis, and Theodoros Tarnanidis

Abstract The visualization abilities of Correspondence analysis (CA) and MCA exploratory data analysis' methods, remains one of their main characteristic and advantage. Yet, there are certain interpretive indicators to be produced and analysed, in order to reach to a certain and solid conclusion. However, many times the final interpretation is not the appropriate one, because a deeper understanding of the methods is required to decide.

In this article, a different approach for the visualization of the methods is proposed and presented. With the contribution of a programming language, R, the three main interpretive indicators of MCA are combined and the appropriate points are selected to be presented. This way, even a non expert in the methods scholar, can be benefited as only the most important points are selected and analysed.

Data from a large social survey (European Social Survey) is used and a specific correspondence between religion practices and attitude towards immigrants is explored.

The proposed method contributes in the exact interpretation of the phenomenon investigated.

Keywords MCA; ESS; interpretive indicators; R language

Vasileios Ismyrlis
University of Macedonia, Greece, email: vasismir@uom.edu.gr

Efstratios Moschidis
University of Macedonia, Greece, email: smos@uom.edu.gr

Theodoros Tarnanidis
University of Macedonia, Greece, email: tarnanidis@uom.edu.gr

Sentiment and return distributions on the German stock market

Emile David Hövel, and Matthias Gehrke

Abstract The systematic forecasting and explanation of stock returns are some of the critical challenges in capital market research. Particularly controversial is the integration of market sentiment into the explanatory models of stock returns. This study shows that sentiment from different sources (survey based, market implicit, and social media-based), if considered as risk factor in linear multifactor models, provides significant explanatory contributions to return distributions on the German stock market.

The observed sentiment premia are negative and weakly correlated with the Carhart factors. Also, a higher rate of α insignificance in models extended by sentiment factors has been ascertained. Besides, it is currently being examined whether innovative classification models such as neural networks are also suitable for systematically verifying corresponding sentiment-induced explanations for returns on the German stock market. First results will be presented in this talk.

Keywords behavioral finance; investor sentiment

Emile David Hövel
Universidad Católica San Antonio de Murcia, Spain, e-mail: edhovel@alu.ucam.edu

Matthias Gehrke
FOM University of Applied Sciences, Germany, e-mail: matthias.gehrke@fom.de

Risk management based on conditional extreme quantile risk measures on energy market

Grażyna Trzpiot, Alicja Ganczarek-Gamrot, and Dominik Krężołek

Abstract The aim of this paper is to describe and measure conditional extreme risk on energy market. The risk was estimated with Conditional Value at Risk (CVaR) and Median Shortfall (MS) base on some types of Value-at-Risk measures: VaR, stress VaR, Incremental Risk Charge (IRC) and Extreme Value Index (EVI). These measures were calculated on time series of daily and hourly rates of return of electric energy prices from the European Energy Exchange (EEX) spot market. Based on time series from 5th October 2008 to 31th December 2016 we attempted to answer the questions: which measure is more appropriate for risk estimation on energy market.

Keywords extreme risk; quantile risk measures; energy market

Grażyna Trzpiot
University of Economics in Katowice, Poland, e-mail: grazyna.trzpiot@ue.katowice.pl

Alicja Ganczarek-Gamrot
University of Economics in Katowice, Poland, e-mail: alicja.ganczarek-gamrot@ue.katowice.pl

Dominik Krężołek
University of Economics in Katowice, Poland, e-mail: dominik.krezolek@ue.katowice.pl

Comparison of systemic risk in the banking sector and selected sectors of real economy – case of Poland

Katarzyna Kuziak, and Krzysztof Piontek

Abstract Banks and other financial companies are less sensitive to the market return than the cyclical construction sector, while they are as sensitive to the market as the non-cyclical food sector. This market sensitivity captures the direct exposure to macro-economic shocks as well as the indirect exposure that arises from contagion as well as feedback effects with the real economy (Muns, BijIsma 2011). This paper compares systemic risk in the banking sector, the construction sector, and the food sector. To measure systemic risk we use econometric measures of connectedness based on principal-components analysis (PCA) and delta Conditional Value at Risk (proposed by Adrian and Brunnermeier, 2011). The CoVaR measure is estimated by employing quantile regressions on weekly rates of return data. In PCA, the rate of increase in principal components can be used as an indicator of systemic risk (Billio et al. 2012, Zheng et al. 2012).

We state following hypothesis: Systemic risk is significantly larger in the banking sector relative to the other sectors. To verify it we need to compare the systemic risk measures for the banking sector with the other non-financial sectors. Empirical research will be conducted for Polish banking sector, the food sector and, the construction sector.

Keywords Systemic risk; banking sector; real economy; CoVaR; PCA

Katarzyna Kuziak
Wrocław University of Economics, Poland, email: katarzyna.kuziak@ue.wroc.pl

Krzysztof Piontek
Wrocław University of Economics, Poland, email: krzysztof.piontek@ue.wroc.pl

Credit risk with credibility theory: a distribution-free estimator for probability of default, value at risk and expected shortfall

Anne Sumpf

Abstract Credibility theory is a distribution-free estimation technique from actuarial science. This paper shows an interpretation of the credibility theory for credit risk and connects it with Bernoulli mixture models. Thus, credibility theory is a generalization of Bernoulli mixture models. Based on credibility theory, we construct distribution-free estimators for probability of default, expected loss, Value-at-Risk and expected shortfall. In the end, the estimators are illustrated by numerical examples.

Keywords Value at Risk; Expected Shortfall; Quantile Credibility

Anne Sumpf
Technische Universität Dresden, Germany, e-mail: anne.sumpf@tu-dresden.de

Flexible clustering

Andrzej Sokołowski, and Małgorzata Markowska

Abstract Flexibility of cluster analysis is sometimes understood as the robustness of final partition of objects to the changes in the list of diagnostic variables – deleting some from the list or adding some. In this paper we propose a procedure which makes possible to calculate distance matrix on the basis of different subsets of variables, but the selection of variables is somehow unified. The procedure starts from the classical standardization of each variable. Before the calculation of a distance between two objects, we eliminate the variable with the biggest absolute value in the first object and in the second object. So we have two variables less in the list. If by chance the same variable is pointed for elimination from both objects, the next variable with the biggest (from both objects) absolute value should be eliminated. With this procedure each element of distance matrix is based on the same number of variables, but variables can be different.

We consider 17 variables describing so called smart society characteristics for 28 European Union countries, as an empirical example. On this data we study the effect of the proposed approach within different distance measures and clustering methods.

Keywords clustering; distance matrix; variable selection

Andrzej Sokołowski
Cracow University of Economics, Poland, e-mail: sokolows@uek.krakow.pl

Małgorzata Markowska
Wrocław University of Economics, Poland, e-mail: malgorzata.markowska@ue.wroc.pl

A coefficient of determination for clusterwise linear regression with mixed-type covariates

Salvatore Ingrassia, and Roberto Di Mari

Abstract The famous coefficient of determination (R^2) is defined as a measure of explained variation over the total observed variation and is universally used to assess the model fit in linear regression. Although extensions to other models are available, they are not as widespread.

We propose a coefficient of determination for clusterwise linear regression that is able to accommodate covariates of any type that extends recent work in cluster analysis. Specifically, starting from a three-term decomposition of the residual sum of squares, we propose an overall R^2 which can be showed to be a weighted sum of local R^2 that individually indicate how good the within-cluster fit is. We illustrate the proposed measure based on a simulation study and an empirical example.

Keywords Mixtures of regressions; Model-based clustering; Model assessment

Salvatore Ingrassia
University of Catania, Italy, email: s.ingrassia@unict.it

Roberto Di Mari
University of Catania, Italy, email: roberto.dimari@unict.it

Two new algorithms, critical distance clustering and gravity center clustering

Farag Kuwil, Radwan Abuissa, Fionn Murtagh, and Umit Atila

Abstract We developed a new algorithm based on Euclidean distance among data points and employing some mathematical statistics operations and called it critical distance clustering (CDC) algorithm (Kuwil, Shaar, Ercan Topcu, & Murtagh, Expert Syst. Appl., 129 (2019) 296–310. https://authors.elsevier.com/a/1YwCc3PiGTBULo). CDC works without the need of specifying parameters a priori, handles outliers properly and provides thorough indicators for clustering validation. Improving on CDC, we are on the verge of building second generation algorithms that are able to handle larger size objects and dimensions dataset.

Our new unpublished Gravity Center Clustering (GCC) algorithm falls under partition clustering and is based on gravity center "GC" and it is a point within cluster and verifies both the connectivity and coherence in determining the affiliation of each point in the dataset and therefore, it can deal with any shape of data and K-means algorithm can be considered as a special case of GCC. In GCC algorithm, lambda is used to determine the threshold and identify the required similarity inside clusters using Euclidean Distance. Moreover, two coefficients lambda and n provide to the observer some flexibility to control over the results dynamically (parameters and coefficients are different, so, in this study, we assume that existing parameters to implement an algorithm as disadvantage or challenge, but existing coefficient to get better results as advantage), where n represents the minimum number of points in each cluster and lambda is utilized to increase or decrease number of clusters. Thus, lambda and n are changed from the default value in case of addressing some challenges such as outliers or overlapping or weakly clusters consisting of a less points compared to others. The basic idea is to find the critical distance lambda from CDC algorithm and MST algorithm. This allows the measurement of ULO where lambda is used to determine a set of effective points are shared at the same threshold, and then choose the lowest edge value as the gravity center for this cluster. It will be made according to some criteria that will be explained in detail through two methods.

Keywords CDC algorithm; clustering; critical distance; gravity center

Farag Kuwil
Karabuk University, Turkey, e-mail: faraghamedali@ogrenci.karabuk.edu.tr, kuwil73@gmail.com

Radwan Abuissa
Ankara Yıldırım-Beyazıt-University, Turkey, e-mail: ra691190@gmail.com

Fionn Murtagh
Huddersfield University, United Kingdom, e-mail: f.murtagh@hud.ac.uk

Umit Atila
Karabuk University, Turkey, e-mail: umitatila@gmail.com

Triplet clustering of one-mode two-way proximities

Akinori Okada, and Satoru Yokoyama

Abstract Most of multidimensional scaling and cluster analysis deal with the relationship of two objects. While the relationship between two objects is important to disclose the proximity structure of objects, sometimes it is not enough just to analyze the relationship of two objects, because the relationship of three or more objects may play an important role in the proximity structure of objects. In the present study, an algorithm of clustering three objects at each step of the clustering is introduced. While the relationship of more than three objects may be important in the proximity structure of objects, it seems reasonable to deal with the relationship of three objects as a first step of dealing with the relationship of more than two objects. There are two types of algorithms of multidimensional scaling and cluster analysis to carry out the analysis of the relationship of three objects. One is to analyze one-mode three-way proximities, and the other is to analyze one-mode two-way proximities. In the present study, an algorithm of cluster analysis of analyzing one-mode two-way proximities, where each step of the clustering is done by agglomerating three objects, is introduced. The algorithm derives an index showing the strength of the relationship of three objects by manipulating three relationships between any two objects from the three, which is one of the two strategies adopted by algorithms for analyzing one-mode three-way proximities by multidimensional scaling and cluster analysis. The problem of this type of algorithm is how to define an index showing the strength of relationship of three objects by manipulating three relationships between any two objects of the three. It is necessary that the resulting index can be convinced that it indicates the strength of the relationship of three objects or how close the relationship of three objects altogether is. The algorithm of the present clustering of the relationship of three objects focusses the attention to the largest and the smallest proximities of the three proximities between any two objects from the three. It is meaningful to develop the algorithm of clustering based on the relationship of three objects by analyzing one-mode two-way proximities, because one-mode two-way proximities prevails more widely than one-mode three-way proximities do, and it is useful to disclose the relationship of three objects underlies one-mode two-way proximities.

Keywords agglomerative; cluster analysis; hierarchical; triadic relationship

Akinori Okada
Rikkyo University, Tokyo, Japan, e-mail: okada@rikkyo.ac.jp

Satoru Yokoyama
Aoyama Gakuin University, Tokyo, Japan, e-mail: yokoyama@busi.aoyama.ac.jp

Societal responsibility of data scientists

Ursula Garczarek and Detlef Steuer

Abstract The biggest, relatively recent changes in practical data science are the availability of vast amount of data together with an increase in computational power. Technically speaking this enables fast, low-cost processing of ever-changing large data bases by algorithms to derive continuously updated, highly condensed and aggregated data that feed into human decision making or are used in rules for automatic decision making.

These possibilities change human interaction and thus society directly and fundamentally. This is most apparent whenever the processed and analysed data are about humans and human behaviour. Therefore data scientists play an important role for this new form of interaction. With this importance comes increased responsibility.

In the talk we want to highlight several ethical issues that can arise from a data scientist's work. This is not meant to be a complete list of all issues but a teaser for discussions. As we will explain, to our perception, data scientists are currently not well equipped to do justice to their responsibility to society in their daily work. One limitation is that there is not really a sense of community among those that perform data science tasks. These people often see themselves not as data scientists, but as engineers, computer scientists, statisticians, mathematicians etc. Secondly, discussions on ethics are rarely seen as being an important part of a professional life, and rarely integrated into scientific education. A sense of community and an environment open to controversial discussions around ethics would foster the development of morality, a set of deeply held, widely shared, and relatively stable values among data science practitioners.

So what can we – scientists of all shades and members of various data science related societies – do about it?

Keywords data science; community; ethics; societal responsibility; education

Ursula Garczarek
Cytel Inc, Clinical Research Services ICC, Switzerland, e-mail: ursula.garczarek@cytel.com

Detlef Steuer
Helmut-Schmidt-Universität, Germany, e-mail: steuer@hsu-hh.de

Data Science Education, Skills and Industry in Europe

Berthold Lausen, Alexander Partner, Stephen Lee, Henrik Nordmark, Mahdi Salhi, and Christopher Saker

Abstract Classification Societies focussed on methodologies as for example classification/supervised learning, clustering/unsupervised learning and multidimensional scaling defining the mathematical foundations of the emerging discipline data science, since the Classification Society was founded in London 1964. The proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996 included data science in the title *'Data Science, Classification, and Related Methods'*. The proceedings of the first European Conference on Data Analysis (ECDA) jointly held by the GfKl Data Science Society and the French-speaking Classification Society (SFC) in July 2013 at the University of Luxembourg demonstrated this further with the title *'Data Science, Learning by Latent Structures, and Knowledge Discovery'*. In 2018 a special issue of the Journal of Data Science and Analytics has papers devoted to current issues in Data Science viewed from a European perspective which were in the European Data Science Conference (EDSC), an invitation only event organised by Professor Sabine Krolak-Schwerdt and her team in November 2016 in Luxembourg as the inaugural conference of the *European Association for Data Science (EuADS)*. In this context we discuss challenges and needs for data science education and skills.

The landscape for school mathematics in the United Kingdom is very different now compared to even just five years ago. Another new compulsory requirement is that students must work with prescribed large data sets during their A level Mathematics studies, and that the use of technology must permeate across their course. During the period of recent changes in the UK school system we have seen the emergence of undergraduate (BSc) and postgraduate taught (MSc) qualifications in Data Science. For example in 2014 the Department of Mathematical Sciences and the School of Computer Science and Electronic Engineering at the University of Essex have introduced a BSc in Data Science and Analytics and an MSc in Data Science. The curriculum of these courses covers compulsory modules from computer science and mathematical sciences, introducing students to a range of mathematics and statistical topics as well as computing skills. As the University of Essex we aim to offer Data Science as a minor subject with major subjects from the faculty of humanities, for example BA Philosophy with Data Science (October 2020 onwards).

In this talk we will also review curricula of data science related university degrees, consider how their content matches industry expectations and discuss plans for further developments.

Keywords Classification Societies; School Education; Data Science Curricula; Industry Expectations

Berthold Lausen
Department of Mathematical Sciences, University of Essex, UK, e-mail: blausen@essex.ac.uk

Alexander Partner
Department of Mathematical Sciences, University of Essex, UK, e-mail: akpart@essex.ac.uk

Christopher Saker
Department of Mathematical Sciences, University of Essex, UK, e-mail: cjsake@essex.ac.uk

Stephen Lee
Mathematics in Education and Industry (MEI), Wiltshire, UK, e-mail: Stephen.lee@mei.org.uk

Henrik Nordmark
Profusion Ltd, London, UK, e-mail: henrikn@profusion.com

Mahdi Salhi
Profusion Ltd, London, UK, e-mail: mahdis@profusion.com

Analysis of statistical tests indications in assessing data conformity to Benford's Law in fraud detection

Józef Pociecha, and Mateusz Baryła

Abstract Benford's Law deals with the probability of the occurrence of significant digits in numbers. One of the applications of the law is to use it as a tool in fraud detection procedures. In the literature, there are three main categories of Benford's Law tests: the primary tests, the advanced tests, and the associated tests. The idea of the primary tests is to verify whether an empirical distribution of digits in numbers conforms to Benford's distribution or not. When assessing data conformity to Benford's distribution, various statistical tests can be employed. The aim of the paper is to compare the indications of the following six statistical tests: chi-square goodness of fit test, Kolmogorov-Smirnov test, tests using distance measures (Euclidean distance, Chebyshev distance), the test based on the means of comparing distributions, and the test based on mantissas. The analysis includes four data sets. Two of them regard domestic and foreign revenues (free of financial frauds) from the sales of finished products of a certain company. The next two data sets, resulting from the conducted experiment involving two strategies of committing frauds, also concern domestic and foreign revenues from the sales of finished products.

Keywords Benford's Law; fraud detection; statistical tests

Józef Pociecha
Cracow University of Economics, Poland, e-mail: jozef.pociecha@uek.krakow.pl

Mateusz Baryła
Cracow University of Economics, Poland, e-mail: mateusz.baryla@uek.krakow.pl

Conditional extreme quantile risk measures on metals market

Dominik Krężołek, and Grażyna Trzpiot

Abstract The paper attempts to assess the risk related to the extreme events, for which the probability of occurrence is very low, while the associated consequences may be catastrophic. Some selected risk models based on extreme quantile measures have been proposed, but considered in the dynamic terms. The Value-at-Risk approach for the high orders quantiles have been used. The accuracy of estimation has been evaluated using conditional VaR models based on Hill's estimator. Empirical analysis was carried out on the metal market

Keywords extreme risk; extreme quantiles; metals market

Dominik Krężołek
University of Economics in Katowice, Poland, e-mail: dominik.krezolek@ue.katowice.pl

Grażyna Trzpiot
University of Economics in Katowice Poland, e-mail: grazyna.trzpiot@ue.katowice.pl

Fuzzy clustering with skew components

Francesca Greselin, Agustín Mayo-Iscar, and Luis Angel Garcìa-Escudero

Abstract Clustering is an important technique in exploratory data analysis, with applications in image processing, object classification, target recognition, data mining etc. and is pervasive in all disciplines. The aim is to partition data according to natural classes present in it, assigning data points that are more similar to the same cluster.

Oftentimes, natural classes are not so evident, and a fuzzy classification is preferable, at least for a subset of the units, where membership values convey the extent of belonging of a data unit to each class.

To achieve a fuzzy clustering that reflects reality, careful attention to data with non-Gaussian shape is of the utmost importance. To date, many proposals for fuzzy clustering are based on C-means, that is the fuzzy version of K-means. The aforementioned methods assume clusters of isotropic shape.

Our proposal is to widen the applicability of fuzzy clustering, by adopting fuzzy skew components. To address this multivariate problem, we choose an objective function related to a mixture of skew normal, and we add fuzziness, through a fuzzifier parameter. We also want to obtain a robust fuzzy classification, as data contamination can completely spoil down non-robust techniques. Robustness is obtained by the joint usage of impartial trimming and constrained parameter estimation. The proposed approach is evaluated by means of simulated as well as empirical data, to show how intermediate membership values are estimated for observations lying at cluster overlap, while cluster cores are composed by observations that are assigned to such cluster in a crisp way.

Keywords Fuzzy clustering; skewness; robust estimation; trimming; constrained estimation; asymmetric cluster shape

Francesca Greselin
University of Milano Bicocca, Italy, francesca.greselin@unimib.it

Agustin Mayo Iscar
University of Valladolid, Spain, agustin@med.uva.es

Luis Angel Garcia Escudero
University of Valladolid, Spain, lagarcia@eio.uva.es

Distance measurement and clustering when fuzzy numbers are used. Survey of selected problems and procedures

Jozef Dziechciarz, and Marta Dziechciarz Duda

Abstract The importance of fuzzy numbers for measurement of socio economic phenomena is widely recognised. The use of fuzzy numbers is thoroughly discussed in the classical literature. The argument is limited to conventional understanding of fuzzy numbers, i.e. symmetric, not overlapping triangular, with equal width. The applicability of such numbers in Computer Aided Web Interviewing is limited. Respondents tend to use asymmetric fuzzy numbers with overlapping shape and unequal width. The use of fuzzy measurement results for multivariate statistical analysis is not straightforward. A number of problems arise, especially when unconventional fuzzy numbers are involved. Primary concept to be defined is the distance and dissimilarity measure. Some proposals may be found in the pattern recognition literature. The empirical tasks involve clustering and classification of fuzzy data along with clustering and classification of fuzzy sets. Sometimes clustering algorithm using both rough and fuzzy sets are to be combined. Most popular approach is distance based similarity measures of fuzzy sets.

Linear ordering of objects or sets is important version of multivariate statistical analysis based on distance and dissimilarity measures calculated with fuzzy numbers.

The analysis of fuzzy clustering and evaluation of results requires procedures for comparison of fuzzy clustering methods.

The empirical verification in households' wellbeing measurement, statistical analysis and econometric modelling will be discussed.

Keywords fuzzy algorithm; fuzzy classification; fuzzy clustering; fuzzy computing; fuzzy data; fuzzy distance; fuzzy heuristics; fuzzy linear ordering; fuzzy measurement; fuzzy numbers; fuzzy pragmatics

Jozef Dziechciarz
Wroclaw University of Economics, Poland, e-mail: Jozef.Dziechciarz@ue.wroc.pl

Marta Dziechciarz Duda
Wroclaw University of Economics, Poland, e-mail: Marta.Dziechciarz@ue.wroc.pl

The impact of the publication of short selling positions on German stock returns

Matthias Gehrke, Jannis Kepper

Abstract In response to the financial crisis of 2007/2008, the European Union (EU) adopted the so-called "Short Sale Regulation" in 2012. The purpose of this regulation is to establish a uniform and transparent regulatory framework for short selling in the EU. A key aspect of the regulation is the publication requirement it introduces for substantial net short selling positions. Previous work about short selling has dealt in particular with the American stock market, as there was no data available for the European market in this granularity in the past. Against this background, the objective of this paper is to find out whether the publication of substantial short positions has a demonstrable impact on stock market returns on the German stock market. To answer the research question, different hypotheses are formulated based on theory and literature review.

First, an event study is performed to determine abnormal returns. Second, a regression analysis is used to explain these returns by using appropriate explanatory variables. The data set used in our studies comprises 6679 events relating to 58 different companies of the German HDAX over a period from November 2012 to December 2017. The results of the event study show that the publication of both the increase and reduction of short selling positions has a significant impact on abnormal returns. In the regression analysis some of the hypotheses where rejected and others were confirmed.

As a result of this work, it should be noted that the market participants who have to publish their short selling positions are well-informed investors who are deliberately generating abnormal returns on the basis of short selling. Furthermore, this work confirms the medium-stringent form of the Efficient Market Hypothesis, according to which the publication of new company announcements, in this case the publication of short selling positions, entails a corresponding capital market reaction. All in all, this paper makes a decisive contribution to the research regarding understanding short selling. The results are not only a contribution to the scientific discussion but can be valuable for investors and regulatory authorities, as it provides insights into the extent to which the adopted transparency regulations influence stock returns.

Keywords short selling; German stock market; short-sale-regulation

Matthias Gehrke
FOM University of Applied Sciences, Germany, e-mail: matthias.gehrke@fom.de

Jannis Kepper
FOM University of Applied Sciences, Germany, e-mail: janniskepper@gmail.com

Japanese women's attitudes towards childrearing: text analysis and multidimensional scaling

Kunihiro Kimura

Abstract I analyzed Japanese women's attitudes towards childrearing by coding topics that appeared in a sample of documents they had written and applying asymmetric multidimensional scaling to the coded data. The sample (n = 102) was taken from the emails published in *Is Childrearing a Strain? Emails from 2,118 Readers*, Special Issue of *AERA*, a Japanese magazine. I used Unilateral Jaccard Similarity Coefficient to measure the similarity between 13 topics and prepared the asymmetric proximity matrix to be analyzed. The result of drift vector model suggested a self-serving bias in causal attribution of success and failure. The emails expressing positive attitudes toward childrearing tended to refer to only personal topics, while those manifesting negative attitudes tended to refer to personal, interpersonal, and societal topics. In the latter group of emails, reference to personal topics such as "captivity" and "relationship with husband" implied reference to interpersonal topics such as "help from others" and "work-life balance," but not *vice versa*. In turn, reference to interpersonal topics implied reference to societal topics such as "schools" and "administrative agencies and policies," but not *vice versa*.

Keywords content analysis; asymmetric proximity; drift vector model; SMACOF

Kunihiro Kimura
Tohoku University, Japan, e-mail: kkimura@m.tohoku.ac.jp

Using domain taxonomy for computational generalization

Boris Mirkin, Dmitry Frolov, Susana Nascimento, and Trevor Fenner

Abstract Consider a domain taxonomy such as the Computing Classification System by the Association for Computing Machinery (ACM-CCS 2012). That is a rooted tree whose nodes are labeled by taxonomy topics. We are concerned with a meaning of term "to generalize" as "to derive or induce (a general conception or principle) from particulars". We consider a fuzzy leaf set S obtained as a result of some process such as clustering. The problem of our concern is finding a most specific generalization of S in the taxonomy. This generalization "lifts" S to its ``head subject", a node in the higher ranks of the taxonomy tree. The head subject is supposed to ``tightly" cover the query set S, possibly bringing in some errors, both ``gaps" and ``offshoots". A gap is a node covered by the head subject but not belonging in S (1st type error). An offshoot is a node in S which is not covered by the head subject (2d type error). Our method ParGenSF globally minimizes a penalty function combining the numbers of head subjects and gaps and offshoots, differently weighted. This method works recursively bottom-up from tree leaves to the root by selecting, at each step, either of two scenarios and corresponding events: (1) the general concept has emerged in the node; or (2) the general concept has not emerged in the node. We apply this to a collection of about 18000 research papers published in 17 Springer journals on Data Science for the past 20 years. We extract a taxonomy of Data Science from ACM-CCS 2012 and augment it by adding a few leaves corresponding to newest approaches. We find fuzzy clusters of leaf topics over the text collection using a text-to-topic relevance score matrix. We apply an additive fuzzy clustering approach to Laplacian normalization of a topic-to-topic co-relevance matrix and derive 6 fuzzy spectral clusters, of which three clusters are especially homogeneous, relating to either Learning or Retrieval or Clustering. We apply the ParGenFS method to generalize each these three clusters. The found head subjects of these thematic clusters are used to comment on the tendencies of current research in the corresponding aspects of the domain. Existing approaches for deriving tendencies of research are based on the analysis of co-citation networks and cannot provide that level of abstraction which is achieved with our approach.

Keywords taxonomy; clustering; generalization; text analysis

Boris Mirkin
NRU Higher School of Economics, Moscow, RF, and Birkbeck University of London, UK, e-mail: bmirkin@hse.ru

Dmitry Frolov
NRU Higher School of Economics, Moscow, RF, e-mail: dfrolov@hse.ru

Susana Nascimento
Universidade Nova de Lisboa, Caparica, Portugal, e-mail: snt@fct.unl.pt

Trevor Fenner
Birkbeck University of London, United Kingdom, e-mail: trevor@dcs.bbk.ac.uk

Detection of topics and time series variation in consumer web communication data

Atsuho Nakayama

Abstract Motivation of this study is to using consumers' uploading habits on internet for marketing purposes. The distribution of sharing photos on social networking sites, instant sharing with smartphones on the internet has been increasing. Today uploading habits have become part of our lives. People use images and text on the internet to represent their personal concerns such as activities, interests and opinions. The desire to identify market trends, the analysis of consumer web communication data has received much attention. We collected Twitter entries about new products based on their specific expressions of personal concerns. We tokenized each tweet message that was written in sentences or sets of words to detect topics more easily. Morphological analyses such as tokenization, stemming, and part-of-speech tagging to separate the words were performed. Next, we selected keywords representative of our chosen topics. We performed a statistical analysis based on the complementary similarity measure that has been widely applied in the area of character recognition. Then, we detected trending topics by classifying words into clusters based on the document-term matrix in web communications among consumers. The document-term matrix was sparse and of high dimensionality, so it was necessary to perform a dimensionality reduction analysis. We must employ some excellent computing resources to help analyze the sparse and large matrix. It's often contained noise, making it difficult to uncover the underlying semantic structure. We analyzed by non-negative matrix factorization as a dimensionality reduction model. Furthermore, personal concerns are influenced by new product strategies, such as marketing communication strategies, and they change over time. It is important to consider the temporal variation of these topics. We clarified temporal variation by using the weight coefficients between entries and topics which obtained from non-negative matrix factorization. The weight coefficients show the strength of associations between entries and topics. We modeled the relationships between entries and topics by Bayesian networks and captured changes in topics over time. Bayesian networks are a type of probabilistic graphical model that builds models from data by using Bayesian inference for probability computations. The causations can be modeled by representing conditional dependence based on edges in a directed graph of Bayesian networks. They are used for a wide range of studies such as prediction, anomaly detection, reasoning, and time series prediction. The new findings obtained by these analyses are reported in this paper.

Keywords hierarchical clustering; error correction

Atsuho Nakayama
Tokyo Metropolitan University, Japan, e-mail: atsuho@tmu.ac.jp

Making product recommendations based on latent topics: an analysis of online purchase data with topic models

Johanna Fischer

Abstract Increasing product variety in online shops make purchase decisions for customers extremely difficult. To address this issue online retailers incorporate recommender systems in their online shops, for example Amazon with the "Customers who bought this item also bought" section or Zalando displaying a set of additional products to "Complete your look". Those recommender systems help customers to find apt products that serve their interests and complete their shopping. They are also vital to companies in enhancing sales and competitiveness. Even though helpful, current recommender models come with various drawbacks. For instance todays recommender systems are prone to recommend products that are frequently bought across customers (leaving out rare products), display products a customer already knows or ignore purchase motivations of customers. A novel approach that addresses those issues are topic models. Topic models were originally developed for application to text data, in particular for finding the latent topics which are present in large text collections. Those models, though, are not necessarily tied to text data but can be applied to various other data sources, including purchase data. A small number of literature has already successfully applied topic models to purchase data for explaining future buying behavior. The present paper extends those current academic works by examining the following three new aspects: (1) Applying topic models and some benchmark models on *sparse purchase data* of an online retailer for recommendation purposes. (2) Using a *sticky topic model* as recommender model and (3) Exploring *different ways of pre-processing* and their influence on product recommendations with topic models. The results show that topic models outperform currently used recommender models when predicting new items to a customer in an online shop characterized by high sparsity. In addition, the right pre-processing to be selected for using topic models as recommender systems is discussed and evidenced by empirical findings.

Keywords recommender systems; topic models; latent Dirichlet allocation; purchase history data

Johanna Fischer
Catholic University Eichstätt-Ingolstadt, Germany, e-mail: johanna.fischer@ku.de

Quantitative analysis of phonological structure used in dialects in Osamu Dazai's works

Naoko Oshiro, Sayaka Irie, and Mingzhe Jin

Abstract In this research, we will clarify the feature of the dialects used in Osamu Dazai's novels from the viewpoint of text analysis. Osamu Dazai (June 19, 1909–June 13, 1948) was a Japanese writer born in the Tsugaru region of Aomori Prefecture, northern Japan. Tsugaru dialect has a different phonological structure than standard Japanese. The pronunciation of map in Tsugaru dialect is "*cuzu*" while in standard Japanese it is "*cizu*". Although most of Dazai's works were written in standard Japanese, unifying the spoken and written styles, some were written using dialects. The purpose of this study is to objectively indicate the differences in dialects between Dazai's works and spoken language. We analyzed and compared datasets which were extracted from Dazai's works, the spoken language of Tsugaru dialects, and standard Japanese. We extracted the phonological features, in particular, the mora, to identify whether Dazai's novels are classified into standard Japanese or dialects. Consonants were replaced with vowels to reflect the features of the Tsugaru dialect, and we analyzed two features of all mora and vowels by statistical methods such as hierarchical clustering. Consequently, compared to the spoken dialects and standard Japanese, Dazai's works which use dialects are similar to standard Japanese. This shows that, regarding Dazai's dialects, unvoiced consonants and phonological structures did not change, for example, like "i" to "u". In Tsugaru dialect, the probability of using "i" is low, but in Dazai's novels, "i" is used in the same way as standard Japanese. Therefore, Dazai's works were written in a dialect close to standard Japanese so anyone could understand them. In this research, we conducted quantitative analysis for Osamu Dazai's works, spoken in Tsugaru dialects and standard Japanese. His works are similar to standard language, rather than dialects, in terms of the structure of phonemes. However, we did not analyze and compare his works written in standard Japanese. For future works, we will examine the features of his standard Japanese by extending the analysis to all his works.

Keywords Osamu Dazai; computational literature; phonological structure; hierarchical clustering

Naoko Oshiro
Graduate School of Culture and Information Science Doshisha University, Japan, e-mail: naoko_oshiro@yahoo.co.jp

Sayaka Irie
Graduate School of Culture and Information Science Doshisha University, Japan, e-mail: sarasara_3ng@hotmail.co.jp

Mingzhe Jin
Faculty of Culture and Information Science Doshisha University, Japan, e-mail: mjin@mail.doshisha.ac.jp

Isotonic boosting procedures for classification

Miguel Fernández, David Conde, Cristina Rueda, and Bonifacio Salvador

Abstract When dealing with classification problems in statistical practice, it is common to have information saying that higher values of some predictors are related to higher (or lower) values of the response. The incorporation of this information to classification procedures yield rules that are easier to interpret and perform better than the ones not accounting for the information. Here, we incorporate this sort of information to boosting procedures developing isotonic boosting procedures, incorporating the information directly in the rule definition through isotonic regression, both for the two-population and the multipopulation cases. We show the good performance of these new procedures not only in simulation but in real data cases coming from different fields.

Keywords Classification; boosting procedures; order restricted inference

Miguel Fernández
Universidad de Valladolid, Spain, e-mail: miguelaf@eio.uva.es

David Conde
Universidad de Valladolid, Spain, e-mail: dconde@eio.uva.es

Cristina Rueda
Universidad de Valladolid, Spain, e-mail: cristina.rueda@uva.es

Bonifacio Salvador
Universidad de Valladolid, Spain, e-mail: bosal@eio.uva.es

Development of indices for the regional comparative analysis of musical compositions, focusing on rhythm

Akihiro Kawase, and Mitsuru Tamatani

Abstract In musicology, style analysis divides musical elements into four parts: sound, harmony, melody, and rhythm, and then examines a work, considering the characteristics of the target composer and target music. According to the qualitative research of Mayer, the works of composers in a similar generation and with similar cultural backgrounds create similar impressions on listeners. Recent studies have sought to elucidate the relationship between music and language, focusing on rhythm from among the four elements of style analysis. Patel and Daniele used nPVI, which they proposed would quantify the difference between stress beats and syllable beats in linguistics, a feature indicating the rhythmic leap of a melody. By comparing the subjects of English and French composers, it was shown that rhythmic leaps are more active in the subjects of English composers than those of French composers. Although style analysis has been carried out in a large number of samples, similarities based on the background factors proposed by Meyer have not been objectively verified. Therefore, in this study, we focused on patterns of pitch length in melodies, and aimed to devise indices that could quantitatively indicate differences in data based on such cultural factors as region and era. In this research, the subjects of 9,798 songs published in *A Dictionary of Musical Themes* were converted to the MusicXML, and a corpus was created in which each datum was labeled with background factors. We generated a vector $\boldsymbol{d}$ having elements of pitch length $\boldsymbol{d}_k$ from each melody in the corpus, and 12 features were calculated based on the nPVI, nPC, and rhythm ratio indices of vector $\boldsymbol{d}$. Using these features as explanatory variables and labels as objective variables, we performed classification experiments using random forests and SVM. Due to space limitation, we described a classification experiment of 4,734 themes by composers from Germany, France, Italy, and England. The discrimination rate measured by LOOCV was 52.69%. Moreover, we clarified that the differences between each country can be classified by the entropy of tone pitch length of each work. In this study, we performed classification experiments based on composers' country of origin and era composition. By developing this method of research, it will be possible to construct a system by which to judge, from a quantitative perspective, the origin of music that has been subjectively evaluated.

Keywords SVM; musical composition; regional classification

Akihiro Kawase
Doshisha University, Japan, e-mail: akawase@mail.doshisha.ac.jp

Mitsuru Tamatani
Doshisha University, Japan, e-mail: mtamatan@mail.doshisha.ac.jp

View selection through meta-learning

Wouter van Loon, Marjolein Fokkema, Botond Szabo, and Mark de Rooij

Abstract Integrating information from different feature sets describing the same set of objects is known as multi-view learning. Such feature sets (views) occur naturally in biomedical sciences as, for example, different types of omics data, different gene sets or genetic pathways, or different medical imaging modalities. Integrating these different sources of information can increase the accuracy of medical classification models. However, collecting biomedical data can be expensive and/or burdening for patients. It is therefore important to reduce the amount of required data collection by selecting only those views that are most important for prediction. The group lasso is an existing method for automatically selecting views as part of the model fitting process, but it is slow to train and, more importantly, tends to select many irrelevant views. We propose an alternative method for this problem based on stacked generalization. The proposed method consists of training an L2-penalized logistic regression model on each view separately, and then training another learning algorithm (a so-called meta-learner) on the predictions of the view-specific models. We show on both simulated and real data that when the (logistic) non-negative lasso is chosen as the meta-learner, the proposed method has a much lower probability of selecting irrelevant views than the group lasso, and is often considerably faster to train, while obtaining a comparable classification performance. Additionally, we investigate how different choices of the meta-learner affect the performance of the proposed method.

Keywords multi-view learning; stacked generalization; feature selection; group lasso

Wouter van Loon
Leiden University, The Netherlands, e-mail: w.s.van.loon@fsw.leidenuniv.nl

Marjolein Fokkema
Leiden University, The Netherlands, e-mail: m.fokkema@fsw.leidenuniv.nl

Botond Szabo
Leiden University, The Netherlands, e-mail: b.t.szabo@math.leidenuniv.nl

Mark de Rooij
Leiden University, The Netherlands, e-mail: rooijm@fsw.leidenuniv.nl

The δ-machine: Classification based on distances towards prototypes

Beibei Yuan, Willem Heiser, and Mark De Rooij

Abstract We introduce the δ-machine, a statistical learning tool for classification based on (dis)similarities between profiles of the observations to profiles of a representation set. In this presentation, we discuss the properties of the δ-machine, investigate the definition of the representation set, and derive variable importance measures and partial dependence plots for the machine. Three choices for constructing the representation set are discussed: the complete training set, a set selected by the clustering algorithm Partitioning Around Medoids (PAM), and a set selected by the *K*-means clustering. After computing the pairwise dissimilarities, these dissimilarities take the role as predictors in penalized logistic regression to build classification rules. This procedure leads to linear classification boundaries in the dissimilarity space, but non-linear classification boundaries in the original predictor space. Moreover, we applied two tailored dissimilarity functions to extend the δ-machine to handle mixed type of predictor variables, the adjusted Euclidean dissimilarity function (AEDF) and the adjusted Gower dissimilarity function (AGDF).

We will apply the δ-machine on two empirical data sets, the Mroz data and the Statlog data. For the Mroz data, we will show the non-linear boundaries in the original predictor space which were derived from the δ-machine. For the Statlog data, we compare the δ-machine with five other classification methods. The results showed that the δ-machine was one of the best methods. Moreover, we will show how the performance of the δ-machine changes by applying different types of the representation set. The obtained results showed that when the δ-machine applied the PAM, the results show a good balance between accuracy and interpretability.

Keywords Dissimilarity space; Nonlinear Classification; the Lasso

Beibei Yuan
Leiden University, Netherlands, email: b.yuan@fsw.leidenuniv.nl

Willem Heiser
Leiden University, Netherlands, email: heiser@fsw.leidenuniv.nl

Mark De Rooij
Leiden University, Netherlands, email: rooijm@fsw.leidenuniv.nl

Tree-base ensemble methods for classification

Daniel Uys

Abstract Ensemble methods combine a large number of simpler base learners to form a collective model that can be used for classification. Learning methods such as bagging and random forests can be regarded as tree-based ensemble methods. In these methods, the standard approach is to express the model as a linear combination of the base learners where the coefficients, associated with the base learners, are all equal, i.e., the base learners are equally weighted. An alternative approach would be to assign to those base learners that are considered important, larger weights.

Within a regression context, this has been done by estimating the coefficients of the base learners using least squares. Since a large number of base learners is involved, the residual sum of squares of the linear combination of base learners has to be penalised by, for example, the lasso penalty. However, the large number of base learners also complicates the minimisation of the coefficients in the penalised residual sum of squares criterion. By using the iterative forward stagewise linear regression algorithm for ensemble methods, estimators of the coefficients of the base learners can be obtained.

In this talk, the principles of the forward stagewise linear regression algorithm are considered within a classification context by assigning different weights to different classification base learners. Various tree-based ensemble methods will be evaluated by applying the weighted classification techniques to simulated, as well as to real life datasets.

Keywords ensemble methods; bagging; random forests

Daniel Uys
Stellenbosch University, South Africa, e-mail: dwu@sun.ac.za

Unsupervised feature selection and big data

Renato Cordeiro de Amorim

Abstract The last decade saw a considerable increase in the availability of data. Unfortunately, this increase was overshadowed by various technical difficulties that arise when analysing large data sets. These include long processing times, large requirements for data storage, and other technical issues related to the analysis of high-dimensional data sets. By consequence, reducing the cardinality of data sets (with minimum information loss) has become of interest to virtually any data scientist. Many feature selection algorithms have been introduced in the literature, however, there are two main issues with these. First, the vast majority of such algorithms require labelled samples to learn from. One should note it is often too expensive to label a meaningful amount of data, particularly when dealing with large data sets. Second, these algorithms were not designed to deal with the volume of data we have nowadays. Here, we introduce a novel unsupervised feature selection algorithm designed specifically to deal with large data sets. Our experiments demonstrate the superiority of our method.

Keywords unsupervised feature selection; clustering; big data

Renato Cordeiro de Amorim
University of Essex, UK, e-mail: r.amorim@essex.ac.uk

A simulation study for the identification of missing data mechanisms using visualisations

Johané Nienkemper-Swanepoel, Niël J le Roux, and Sugnet Gardner-Lubbe

Abstract Understanding the cause of the missingness in data is a science of its own and is of great importance for the application of valid and unbiased analysis techniques for missing data. The distribution of missingness is defined by certain dependencies on either observed or missing values in a data set and therefore requires a multivariate visualisation to attempt to identify the missing data mechanism (MDM). Multivariate categorical data sets containing missing data entries can be separated into observed and unobserved (or missing) subsets by applying what is known as active handling of missing values when constructing the indicator matrix. Single active handling of a missing categorical response requires creating an additional category level (CL) for each variable with missing responses. Subset multiple correspondence analysis (sMCA) can then be applied to the recoded indicator matrix to obtain a biplot for the observed subset and a biplot for the missing subset separately. The sMCA biplot of missing categories enables the exploration of the missing values which could expose discerning patterns. It is hypothesised that if insufficient clustering structures occur between the missing CLs in the sMCA biplot of missing values, this could indicate independence and therefore relate to the missing completely at random (MCAR) MDM. The MCAR MDM assumes that the missing values occur independently. The contrary is hypothesised for sufficient clustering structures, which could indicate missing CLs that are highly associated and therefore reflect a missing at random (MAR) MDM. Partitioning around medoids (pam) clustering is used to identify the MDM in different simulated scenarios. A simulation study consisting of data sets with different sample sizes are generated from three distributions. Artificial missingness is created by deleting values according to MAR and MCAR MDMs with different percentages of missing values. The influence of the underlying distribution on the outcome of the clustering techniques will be presented. The insight obtained from the simulation results will be applied to identify the MDM in a real application.

Keywords categorical data; clustering; missing data; missing data mechanism; subset multiple correspondence analysis

Johané Nienkemper-Swanepoel
Stellenbosch University, South Africa, e-mail: nienkemperj@sun.ac.za

Niël J le Roux
Stellenbosch University, South Africa, e-mail: njlr@sun.ac.za

Sugnet Gardner-Lubbe
Stellenbosch University, South Africa, e-mail: slubbe@sun.ac.za

Functional linear discriminant analysis for several functions and more than two groups

Sugnet Lubbe

Abstract Canonical variate analysis in a multivariate data analysis setting aims to find a linear combination of variables which, after transformation to the canonical space, are optimally separated among groups. First focussing on the two-group case, a single canonical variate is defined maximising the between group relative to within group variance ratio. Many functional data analysis methods are based on multivariate data methods where instead of dealing with variables, continuous functions are a generalisation with. Functional linear discriminant analysis as such a generalisation have been discussed by several authors in the literature. Another point of view is to have continuous functions and to search for a linear combination of the functions, such that the resulting functions are optimally separated in the canonical function space.

In this paper a new suggestion for the latter problem is proposed. Two possible solutions are evaluated – finding a single set of coefficients to perform the canonical transformation, or finding time-varying coefficients, i.e. functions of coefficients to transform each time point to a canonical functional space. Furthermore, both these methods can be generalised to discriminant analysis for groups. An optimal two- or three-dimensional visualisation of the canonical functional space is constructed and illustrated with an example.

Keywords linear discriminant analysis; functional data analysis; classification

Sugnet Lubbe
Stellenbosch University, South Africa, e-mail: slubbe@sun.ac.za

Visualising Multivariate Data in a Principal Surface Biplot

Raeesa Ganey, and Sugnet Lubbe

Abstract Biplots are considered as extensions of the ordinary scatterplot by providing for more than three variables. The PCA biplot is one of the basic biplots, that uses the singular value decomposition as an approximation to a data set, where it is then considered to be the biplot plane. A biplot is predictive if information on variables are added in such a way that it allows the values of the variables to be estimated for points in the biplot. Prediction trajectories are created on the biplot to allow information about variables to be estimated. The goal is to extend the idea of biplot methodology by replacing the biplot plane by a principal surface. A principal surface is a smooth two-dimensional surface that passes through the middle of a p-dimensional data set. The distance from the data points to the principal surface are minimized, thus providing a non-linear summary of the data. The formation of a surface is found using an iterative procedure which starts with a linear summary. Each successive iteration is a local average of the points, where an average is based on a projection of a point onto the surface of the previous iteration. The emphasising is on high dimensional data where the biplot based on a principal surface allows for visualisation of samples and the predictive variable trajectories. In this talk, I will compare predictions of samples using a principal surface biplot compared to a PCA biplot for non-linear data.

Keywords Biplots; Principal surfaces; Principal component analysis; Multidimensional scaling

Raeesa Ganey
University of Witwatersrand, South Africa, e-mail: Raeesa.ganey@wits.ac.za

Sugnet Lubbe
Stellenbosch University, South Africa, e-mail: slubbe@sun.ac.za

Using separate sampling to understand mobile phone security compliance

Rénette Blignaut, Isabella Venter, and Humphrey Brydon

Abstract In this study separate sampling was applied to various modelling procedures to assist in the identification of the most important variables describing mobile phone users who are security compliant. The data used in this study combined four study cohorts where the same study design and questionnaire was used. After initial analysis of the data (n=448) it was found that only 7% of mobile phone users, reported applying security measures to protect their phones and/or their personal information stored on their devices. As the proportion of the target classifier was so small, predictive modelling procedures failed to produce accurate models. Separate sampling proportions were introduced to establish if classification accuracy could be improved. This study tested prior class probabilities of 0.3, 0.4, 0.5, 0.6 and 0.7 and assessed models fit using the original data where no separate sampling was applied. Models included: decision trees, 5-fold cross-validation decision trees, logistic regression, neural networks and gradient boosted decision trees. A bagging technique was also introduced for some of the modelling procedures. The results showed that the decision tree model using 0.5 as prior probability produced the most accurate and informative model with a 12.1% misclassification rate. Variables influencing mobile security compliance included age, gender and various security/privacy related behaviors.

Keywords separate sampling; decision tree; neural network; boosting; bagging; logistic regression; prior probability

Rénette Blignaut
University of the Western Cape, South Africa, e-mail: rblignaut@uwc.ac.za

Isabella Venter
University of the Western Cape, South Africa, e-mail: iventer@uwc.ac.za

Humphrey Brydon
University of the Western Cape, South Africa, e-mail: hbrydon@uwc.ac.za

Model based clustering through copulas: parsimonious models for mixed mode data

Dimitris Karlis, Fotini Panagou, and Ioannis Kosmidis

Abstract In a recent paper Kosmidis and Karlis (2016) proposed model based clustering based on multivariate distributions defined through copulas. This approach offers a number of advantages over existing methods mainly due to the flexibility to define appropriate models in certain different circumstances. In this talk we exploit the ideas of extending the approach for higher dimensions. The central idea is to use a Gaussian copula and implement the correlation matrix of the Gaussian copula through certain parsimonious representations giving rise to models of different complexity. We use two different approaches, the first makes use of factor analyzers based on the factor decomposition of the correlation matrix and the second is based on Choleski type decompositions. Application with real and simulated data will be also described.

Keywords: mixture models; model based clustering; mixed mode data

Dimitris Karlis
Athens University of Economics and Business, Greece, email: karlis@aueb.gr

Fotini Panagou
Athens University of Economics and Business, Greece, email: fwtpanagou@aueb.gr

Ioannis Kosmidis
University of Warwick, United Kingdom, email: ioannis.kosmidis@warwick.ac.uk

Clustering ranked data using copulas

Marta Nai Ruscone

Abstract Clustering of ranking data aims at the identification of groups of subjects with a homogenous, common, preference behavior. Human beings naturally tend to rank objects in the everyday life such as shops, one's place of living, choice of occupations, singers and football teams, according to their preferences. More generally, ranking data occurs when a number of subjects are asked to rank a list of objects according to their personal preference order. The input in cluster analysis is a dissimilarity matrix quantifying the differences between rankings of two subjects. The choice of the dissimilarity dramatically affects the classification outcome and therefore the computation of an appropriate dissimilarity matrix is an issue. Several distance measures have been proposed for ranking data. We propose generalizations of this kind of distance using copulas adapted to the case of missing data. We consider the case of the extreme list where only the top-k and/or bottom-k ranks are known. We discuss an optimistic and a pessimistic imputation of missing values and show its effect on the classification. Those generalizations provide a more flexible instrument to model different types of data dependence structures and consider different situations in the classification process. Simulated and real data are used to illustrate the performance and the importance of our proposal.

Keywords copulas; ranked data; dissimilarity; cluster analysis

Marta Nai Ruscone
LIUC Università Cattaneo, Italy, e-mail: mnairuscone@liuc.it

Linking different kinds of omics data through a model-based clustering approach

Vincent Vandewalle, Camille Ternynck, and Guillemette Marot

Abstract In this work, a mixture model allowing for genes clustering using both microarray (continuous) and RNAseq (count) expression data is proposed. More generally, it answers the clustering of variables issue, when variables are of different kinds (continuous and discrete here). Variables describing the same gene are constrained to belong to the same cluster. This constraint allows us to obtain a model that links the microarray and RNAseq measurements without needing parametric constraints on the form of this link. The proposed approach is illustrated on simulated data, as well as on real data from TCGA (The Cancer Genome Atlas).

Keywords clustering; mixed-type data; omics data; mixture models

Vincent Vandewalle
Université de Lille & Inria, France, e-mail: vincent.vandewalle@inria.fr

Camille Ternynck
Université de Lille & Inria, France, e-mail: camille.ternynck@univ-lille.fr

Guillemette Marot
Université de Lille & Inria, France, e-mail: guillemette.marot@inria.fr

A probabilistic distance algorithm for nominal data

Francesco Palumbo, Mario Migliaccio, and Cristina Tortora

Abstract In the cluster analysis framework among the partitioning algorithms, the probabilistic distance (PD)-clustering is an iterative, distribution free, probabilistic clustering method that assigns units to a cluster according to their probability of membership, under the assumption that the product of the probability and the distance of each point to any cluster center is a constant. PD-clustering is more flexible that the classical k-means, but its performance are strongly affected by the whole dimensionality. When the total number of variables increases PD-clustering becomes unstable and may not converge to the solution. More recently, an enhanced version of the algorithm, namely factor PD-clustering (FPDC) was proposed. It integrates PD-clustering and dimensionality reduction into one iterative, alternate-steps procedure for high dimensional data.

This proposal presents a generalization of both the PD-clustering and FPDC algorithms to nominal variables.

Dealing with nominal variables, Euclidean distance geometric properties do not hold anymore. Let p be the total number of variables and Q their corresponding categories, being n the number of units, in a complete disjunctive coding data are arranged in a $n \times Q$ binary data. Under this formalization, the Chi-square distance can be computed between each statistical unit and the K cluster average profiles. Where K states for the a priori known number of the clusters. The proposal shows that the PD-clustering based on the Chi-square distance holds its geometric properties and performs well as long as the global Q dimensionality remains reasonably low. However, tending Q to be large, this contribution demonstrates that the FPDC can also be generalized to the multivariate nominal case, and that it ensures comparable and better results with related approaches, under some specific conditions.

Keywords probabilistic distance clustering; categorical data; dimensional reduction

Francesco Palumbo
University of Naples Federico II, Italy, e-mail: fpalumbo@unina.it

Mario Migliaccio
University of Naples Federico II, Italy, e-mail: migliaccio.mario@gmail.com

Cristina Tortora
San Jose State University, CA USA, e-mail: cristina.tortora@sjsu.edu

Stability of joint dimension reduction and clustering

Michel Van De Velden, Angelos Markos, and Alfonso Iodice D'enza

Abstract Several methods for joint dimension reduction and cluster analysis of categorical, continuous or mixed-type data have been proposed over time. These methods combine dimension reduction (PCA/MCA/PCAmix) with partitioning clustering (K-means) by optimising a single objective function. Cluster stability assessment is a critical and inadequately discussed topic in the context of joint dimension reduction and clustering. Resampling methods provide an elegant framework to assess the stability of complete partitions or single clusters. In this work, we present a resampling scheme that combines bootstrapping and a measure of cluster agreement to assess global cluster stability of joint dimension reduction and clustering solutions and a Jaccard similarity approach for empirical evaluation of stability of individual clusters. Assessment of cluster stability can be also used to inform the selection of the correct number of clusters and dimensions, two parameters that have to be determined in advance. These approaches are implemented in the R package clustrd.

Keywords cluster analysis; dimension reduction; bootstrapping; parameter selection; stability

Michel Van De Velden
Erasmus University Rotterdam, The Netherlands, email: vandevelden@ese.eur.nl

Angelos Markos
University of Thrace, Greece, email: amarkos@gmail.com

Alfonso Iodice D'enza
University of Naples, Frederico II, Italy, email: iodicede@gmail.com

Hierarchical clustering through a penalized within-cluster sum-of-squares criterion

Patrick J.F. Groenen, Yoshikazu Terada, and Mariko Takagishi

Abstract Classical hierarchical clustering algorithms such as single, average, and complete linkage are often based on heuristics without explicitly minimizing a loss function. The Ward clustering approach is an agglomerative hierarchical clustering method that joins those two clusters that minimize the increase in within-cluster sum-of-squares. This method can be seen as an algorithm that yields a hierarchy of nested clusters optimizing a criterion.

In this paper, we propose a novel loss function that combines the within-cluster sum-of-squares with a penalty term that consists of the sum of the qth root of the cardinality of the clusters, where $0 < q \leq 1$. The effect of this penalty term is that for a larger penalty strength parameter λ it is beneficial to merge two or more clusters. The smaller the q, the larger the penalty on small clusters and, thus, this approach aims at removing small clusters quickly. It turns out that that the complete path of λs can be found in n steps simultaneously with the clusters to be merged (with n the number of objects to be clustered). Ward clustering is a special case with $q = 1$. We compare the dendrograms of our new method with other approaches.

Keywords Clustering; hierachical clustering; loss function; regularization

Patrick J.F. Groenen
Econometric Institute, Netherlands, email: groenen@ese.eur.nl

Yoshikazu Terada,
Osaka University, Japan, email: terada@sigmath.es.osaka-u.ac.jp

Mariko Takagishi
Osaka University, Japan, email: takagishi@sigmath.es.osaka-u.ac.jp

PerioClust: a new Hierarchical Agglomerative Clustering method including temporal ordering constraints

Lise Bellanger, Arthur Coulon, and Philippe Husi

Abstract Constrained clustering is a class of semi-supervised learning algorithms. It differs from its unconstrained counterpart by integrating previous knowledge on data to clustering algorithms. In this work, we propose a new Hierarchical Agglomerative Clustering (HAC) procedure including temporal ordering constraints. It is designed to consider two potentially error-prone sources of information associated with the same observations. These sources are of different natures, one quantitative or qualitative and the other reflecting the temporal adjacency structure defined as a binary matrix of connectivity among observations. A distance-based approach is adopted to modify the distance measure in the classical HAC algorithm. The new distance is built using a convex combination of the dissimilarity matrices associated with the two sources of information. Since errors have no reason to be of the same order of magnitude, the two sources must be merged in a balanced way. The choice of the mixing parameter is therefore the key point. We define a selection criterion for this parameter based on the absolute difference between two cophenetic correlations, as well as a resampling procedure to ensure the good robustness of the proposed clustering method. We illustrate it with archaeological data from the prestigious Angkor site in Cambodia.

Keywords semi-supervised learning algorithm; hierarchical clustering; Constrained clustering; data analysis in archaeology

Lise Bellanger
Université de Nantes, France, e-mail: lise.bellanger@univ-nantes.fr

Arthur Coulon
Université de Nantes et Université de Tours, France, e-mail: arthur.coulon@univ-nantes.fr

Philippe Husi
CNRS/ Université de Tours, France, e-mail: philippe.husi@univ-tours.fr

Iterated dissimilarities and some applications

François Bavaud

Abstract A dissimilarity is squared Euclidean iff the eigenvalues of the (weighted) scalar products matrix are non negative. Replacing the latter matrix by its square defines a new dissimilarity, which is squared Euclidean by construction, we refer to as the *iterated dissimilarity*.

The iterated dissimilarity enjoys a simple and direct expression in terms of the original dissimilarities and the object weights, without needing to recourse to spectral decomposition. In particular, one-dimensional configurations are, up to a multiplicative constant, left invariant by iteration.

Iterated dissimilarities turn out to help analysing the relationship between variants of stress minimization on one hand, and p.s.d. approximations of scalar products in the Frobenius norm on the other hand.

They also open presumably new and promising perspectives for the imputation of missing values, in particular for nearly one-dimensional configurations, such as encontered in seriation problems in archeology or psychology.

This contribution details in particular the analysis of dissimilarities between members of the Swiss National Council in the present legislative period, based upon their votes. Direct comparisons are not feasible in view of the turnover among the members, some of them seating during disjoint periods, but recourse to the iterated dissimilarity provides a way of estimating all dissimilarity pairs. The resulting MDS configuration consists of a near-perfect horseshoe, reflecting a one-dimensional left-right ordering of the Council members and their political parties.

Keywords missing values; multidimensional scaling; ordering; seriation; stress

François Bavaud
University of Lausanne, Switzerland, e-mail: fbavaud@unil.ch

Constrained three-way clustering around latent variables approach

Véronique Cariou, and Tom F. Wilderjans

Abstract The proliferation of data, which occurs at an ever-increasing rate, leads researchers to analyze more complex data structures such as three-way arrays where several variables are measured on a set of objects at different occasions or alternatively by different subjects. For example, in sensometrics, rapid profiling techniques like check-all-that-apply and free sorting procedures naturally generate three-way datasets and one important issue remains to detect potential segments. For this purpose, a clustering around latent variables approach, CLV3W, has been proposed for the analysis of three-way data. CLV3W groups the variables (belonging to the second mode) into Q clusters such that the variables within each cluster are as much related (i.e., highest squared covariance) as possible with the associated latent component. An extra feature of this strategy is that a system of weights is associated with each occasion (corresponding to the third mode) that reflects the degree of agreement of that occasion with the latent component from each cluster. In this presentation, a constrained weighting system is proposed in which each cluster shares the same weights among the occasions. This approach is illustrated on a case study pertaining to sensory evaluation.

Keywords clustering of variables; CANDECOMP/PARAFAC; three-way data; sensometrics; constrained CLV3W

Véronique Cariou
StatSC, ONIRIS, INRA, France, e-mail: veronique.cariou@oniris-nantes.fr

Tom F. Wilderjans
Institute of Psychology, Leiden University, Netherlands, e-mail: t.f.wilderjans@fsw.leidenuniv.nl

Clustering binary data by application of combinatorial optimization heuristics

Javier Trejos, Luis Amaya, Alejandra Jiménez, Alex Murillo, Eduardo Piza, and Mario Villalobos

Abstract We study clustering methods for binary, or 0/1, data. This kind of data require the definition of aggregation criteria that measure the compactness of each cluster; we study some of these criteria and deduce some theoretical properties. Five new and original methods for clustering binary data are introduced, using neighborhoods and population behavior combinatorial optimization metaheuristics: first ones are simulated annealing, threshold accepting and tabu search, and the others are a genetic algorithm and ant colony optimization. The methods are compared to classical ones, such as hierarchical clustering, and two versions of k-means: dynamical clusters and partitioning around medoids or PAM. The methods are implemented, performing the proper calibration parameters in the case of heuristics, to ensure good results. From a set of 16 data tables generated by a quasi-Monte Carlo experiment, a comparison of the results obtained by classifying the objects in each data table with the different methods is performed. Furthermore, the comparison of results using two measures of dissimilarities (Jaccard and L_1) is done, and the use of two types of aggregations. Generally speaking, heuristics perform very well, especially compared to classical methods.

Keywords clustering; binary data; simulated annealing; threshold accepting; tabu search; genetic algorithm; ant colony optimization

Javier Trejos
University of Costa Rica, Costa Rica, e-mail: javier.trejos@ucr.ac.cr

Luis Amaya
University of Costa Rica, Costa Rica, e-mail: solomandalo@gmail.com

Alejandra Jiménez-Romero
Costa Rica Institute of Technology, Costa Rica, e-mail: alejimenezr@gmail.com

Alex Murillo
University of Costa Rica, Costa Rica, e-mail: alex.murillo@ucr.ac.cr

Eduardo Piza
University of Costa Rica, Costa Rica, e-mail: eduardojpiza@gmail.com

Mario Villalobos
University of Costa Rica, Costa Rica, Costa Rica, e-mail: mario.villalobos@ucr.ac.cr

Testing for equation of distance-based regressions to see whether two groups form a species

Christian Hennig, and Bernhard Hausdorf

Abstract Biological species are characterized by genetic exchange, which leads to genetic similarity. But groups of individuals that are geographically distant can be genetically more different even if they belong to the same species. It is therefore difficult to decide whether such groups belong to the same species. This issue is treated here by testing whether genetic distances and geographical distances are related within the groups in the same way as between the groups, which would be compatible with them belonging to the same species. Regressions of genetic distance dependent on geographical distance within and between the groups are compared using the classical jackknife principle, taking into account the dependence between distances involving the same individual. In this way we avoid distorting the within- and between-species distributions of geographical distances, which would happen using approaches such as permutation, partial Mantel or bootstrap tests.

Keywords species delimitation; distance-based regression; jackknife

Christian Hennig
University of Bologna, Italy, e-mail: christian.hennig@unibo.it

Bernhard Hausdorf
University of Hamburg, Germany, e-mail: fb5a071@uni-hamburg.de

Mental health: analytical focus and contextualization for deriving mental capital

Fionn Murtagh

Abstract The contextualizing of large and complex data sources is crucial for health and many other situations. Associated with analytical focus can be the addressing of bias in social media and other data sources, and associated with contextualization is Big Data calibration.

In this work, our main objective is the evaluation of national mental health survey data. While specific findings and outcomes are the major objectives here, relating to definition and properties of mental capital, and a further objective is as follows: to plan with metadata and ontology for further, future and rewarding integration with other data sources, both nationally and globally.

An important analytical issue is the resolution scale of the data. A further development in methodology is the clustering of all that is exceptional and anomalous, counterposed to commonality, in Big Data analytics. An important role in the analytics here is to have the data re-encoded, such as using p-adic data encoding, rather than real-valued data encoding. For text mining, and also for medical and health analytics, the analysis determines a divisive, ternary (i.e. p-adic where $p = 3$) hierarchical clustering from factor space mapping. Hence the topology (i.e. ultrametric topology, here using a ternary hierarchical clustering), related to the geometry of the data (i.e. the Euclidean metric endowed factor space, semantic mapping, of the data, from Correspondence Analysis). Determined is the differentiation in Big Data analytics of what is both exceptional and quite unique relative to what is both common and shared, and predominant.

Keywords correspondence analysis; contextualization; analytical resolution scale; hierarchical clustering – ultrametric topology

Fionn Murtagh
Huddersfield University, UK, email: fmurtagh@acm.org

A deep learning analytics to detect prognosis of HCC

Taerim Lee

Abstract Deep Learning Analytics uses predictive models that provide actionable information for HCC patient's better prognosis. It is a multidisciplinary approach based on HCC data processing, AI technology-learning enhancement, HCC data mining, and visualization. Three key components need further clarification to help them effectively apply deep learning in HCC prognosis to explain the methods for conducting deep learning, the benefits of using deep learning and the challenges of using learning analytics in HCC. Discover significant clinical factor and SNP markers to detect prognosis of HCC. Compare the efficiency with other prognosis model using support vector machine, liner discriminant, random forest, logistic regression by ROC curves.

Using ICD-9 codes for HCC, 965 patients with HCC and all available data variables required to develop and test models were identified from a clinical and SNP records database. Data on 645 patients was utilized for development of the model and on 320 patients utilized to perform comparative analysis of the models. Clinical data such as presenting signs & symptoms, socio demographic data, presence of metastatsis, laboratory data and corresponding diagnosis and outcomes were collected. Clinical data and SNP collected for each patient was utilized by to retrospectively ascertain optimal management for each patient. Clinical presentations and corresponding treatment was utilized as training examples.

Keywords deep learning; prognosis

Taerim Lee
Korea National Open University, e-mail: trlee@knou.ac.kr

Analysis of the regional difference of number of patients with blood coagulation disorders in Japan

Shinobu Tatsunami, Kagehiro Amano, Akira Shirahata, and Masashi Taki

Abstract Number of Japanese patients with coagulation disorders has been renewed every year by the Nationwide Survey on Coagulation Disorders. The subtotals of patients in each of four disease groups are summarized in the annual report of the survey, in which changes in the subtotals with respect to Japanese geographical regions are also provided. However, the difference among regions has not been evaluated precisely. Thus the changes in the distribution pattern of the patients' numbers of each disease groups depending on the regions were analyzed in the contingency table. Regions in Japan were numbered from 1 to 10, almost according to the direction from north to south. Coagulation disorders were classified into four groups as follows: 1, hemophilia A; 2, hemophilia B; 3, von Willebrand disease; 4, other coagulation disorders. Correspondence analysis was performed by using the numbers of the region and disease group as explanatory and objective variables of the contingency table. The reported numbers of patients dated at the end of May 2018 were used as the frequency element of the table. Regarding the distribution pattern of the number of patient in four disease groups, there were very similar regions and some different ones. Regarding the disease group, hemophilia A and hemophilia B located very near positions on the biplot (Component 1 and 2), while, von Willebrand disease and other coagulation disorders located independent positions that are different from the positions of hemophilia A and hemophilia B. The correspondence analysis provided a clear classification of the distribution patterns of the patients' numbers for both region and disease. The regions from 1 to 10 sometimes correspond to units in which local medical plans are considered. Therefore, figuring out the regional characteristics will be meaningful.

Keywords correspondence analysis; contingency table; hemophilia

Shinobu Tatsunami
Department of Pediatrics, St. Marianna University School of Medicine, Japan, e-mail: s2tatsu@marianna-u.ac.jp

Kagehiro Amano
Department of Laboratory Medicine, Tokyo Medical University Hospital, Japan, e-mail: kamano@tokyo-med.ac.jp

Akira Shirahata
Kitakyushu Yahata-Higashi Hospital, Japan, e-mail: a-shirahata@kitakyu-hp.or.jp

Masashi Taki
Department of Pediatrics, Yokohama City Seibu Hospital, St. Marianna University School of Medicine, Japan, e-mail: m2taki@marianna-u.ac.jp

Analysis of the power balance of the companies of the "keirersu" with the asymmetric MDS

Tadashi Imaizumi

Abstract In Japanese stock markets, the "keiretsu" system has been a long-standing tradition of companies typically with shareholdings. The members companies of the "keiretsu" manage small portions of the shares in each other's companies so-called "cross-holdings". This system absorbs the stock market fluctuations on the member companies and the member companies are able to plan the long-term projects in the past.

Though these relationships and shareholding has been weakened by several reasons, for example, the requirement by Japanese Government and the difficulty of building real momentum etc., this shareholdings system has been undertaken. Japanese Major Banks try to keep this cross-holdings. Each member company will decide on how to keep the stocks of the other member companies strategically. We try to reveal the power balance among the member companies of the "keiretsu" in Japanese stock markets from the matrix of the cross-holdings stocks. The asymmetric multidimensional scaling model with the ellipsoid radius (Okada 1990), is applied to this matrix to reveal the hidden power balance of the member companies.

Keywords asymmetric dimensions; ellipse model; asymmetric proximity; non-metric scaling

Tadashi Imaizumi
Tama University, Japan, email: imaizumi@tama.ac.jp

A fast electric vehicle planner using clustering

Jaël Champagne Gareau, Éric Beaudry, and Vladimir Makarenkov

Abstract In the last few years, several studies have considered the Electric Vehicle Path Planning with intermediate recharge (EVPP-R) problem, that consists of finding the shortest path (according to time) between two points by traveling between pairs of charging stations, while respecting the range of the vehicle. Unfortunately, the exact solution to this problem has a high computational cost. Therefore, speedup techniques are generally necessary (e.g., contraction hierarchies). In this paper, we test a new graph clustering technique on a real map with charging stations data. We show that by regrouping the charging stations that are close to each other into clusters, we can reduce the number of stations considered by a factor of 8 and increase the speed of computation by a factor of 35, while having a very limited tradeoff of less than 1% increase on the average journey duration time.

Keywords electric vehicles; charging stations; planning; clustering; graphs

Jaël Champagne Gareau
Université du Québec à Montréal, Canada, e-mail: champagne_gareau.jael@courrier.uqam.ca

Éric Beaudry
Université du Québec à Montréal, Canada, e-mail: beaudry.eric@uqam.ca

Vladimir Makarenkov
Université du Québec à Montréal, Canada, e-mail: makarenkov.vladimir@uqam.ca

The technology innovation and the critical raw material stock

Beatrix Margit Varga, and Kitti Fodor

Abstract We live in a dynamically changing world. There were times when the use of cars might have been unimaginable, but today the news are about self-driving cars, and the electric cars are more and more popular. Or if our example is the communication, 150 years ago the phone or mobile phone was unknown, but today we have the whole world in our pocket. There were so many innovations in the last few years that raw materials became really indispensable.

The European Commission collected the information in separated studies for the critical raw materials (CRM) and for the non-critical raw materials. This paper is based on this information. However, the development of technology rewrote the list of the currently important raw materials. The European Commission's first list of critical raw materials was completed in 2011. In 2011 there were14 materials on this list, but in 2017 this list contained already 27 materials.

Critical raw materials play a key role in technological innovation; they are the necessary raw materials for many innovations. Critical raw materials are raw materials that are economically and strategically important, but they have high supply risk and are difficult to substitute with other materials.

Such raw materials are germanium, helium, magnesium or cobalt and graphite, which are necessary for many innovations.

In this paper our aim is to identify groups using hierarchical cluster analysis and to identify which clusters are important for innovation. We selected three variables for cluster analysis. These variables are EI (economic importance), SR (supply risk) and „end of life recycling input rate". Our database includes 73 raw materials and we identified 5 homogene groups. There is a group, which seems really important, because it includes only critical raw materials.

Keywords statistics; critical raw material; cluster analysis

Beatrix Varga
University of Miskolc, Hungary, e-mail: stbea@uni-miskolc.hu

Kitti Fodor
University of Miskolc, Hungary, e-mail: f.kitty0408@gmail.com

Knowledge graph mining and affinity analysis for product recommendation on online-marketplace platforms

Siti Nur Muninggar, Reza Aditya Permadi, Simon Simbolon, Verra Mukty, and Putri Wikie Novianti

Abstract This paper investigates application of recommender systems to provide complementary items on an unstructured customer-to-customer e-commerce use case. As opposed to structured marketplace where each item is identified by its global trade item number, our use case presents a particular challenge because taxonomy between similar products is rather limited, where it could only be grouped at most to category level (e.g. drinks). Consequently, the system may recognize similar products as different entities. Market basket analysis (MBA) is one of the most common technique to solve complementary product recommendation problem. However, due to the large number of similar items sold by different sellers, this translates into a high dimensional sparse matrix, which makes finding an effective set of frequently bought together products by using classical MBA is computationally expensive. As one of the alternatives to solve this problem, we apply a multi-stage machine learning technique consisting of knowledge graph mining, clustering and affinity analysis. From a large number of user sessions, a knowledge graph is formed, where products are represented as nodes, and how many times two products are interacted together in a session as the weight of the edges. Furthermore, we applied speaker-listener label propagation algorithm to cluster the graph. Once the cluster is formed among the nodes, active products are mapped to their corresponding cluster, which dramatically reducing the size of the entities by grouping similar products together. We performed affinity analysis on user transaction data with the cluster mapping to find the most frequently bought together entities. We found fifty-three clear-cut association rules, which could not previously be detected by classical affinity analysis. We brought the results to our live recommendation system and it contributed to the growth of the click and paid rate by 4% and 9%, respectively. Our approach was proven to give more relevant and more accurate products to the buyers.

Keywords knowledge graph; market basket analysis; affinity analysis; speaker-listener label propagation; clustering

Siti Nur Muninggar
PT Bukalapak.com, Indonesia, e-mail: siti.muninggar@bukalapak.com

Reza Aditya Permadi
PT Bukalapak.com, Indonesia, e-email: reza.permadi@bukalapak.com

Simon Simbolon
PT Bukalapak.com, Indonesia, e-email: simon.simbolon@bukalapak.com

Verra Mukty
PT Bukalapak.com, Indonesia, e-email: verra.mukty@bukalapak.com

Putri Wikie Novianti
PT Bukalapak.com, Indonesia, e-mail: putri.wikie@bukalapak.com

Pension expenditure modelling and classification analysis

Kimon Ntotsis, Marianna Papamichail, Peter Hatzopoulos, and Alex Karagrigoriou

Abstract The purpose of this work is the modelling of Public Pension Expenditures as percentage of Gross Domestic Product (GDP) of various European countries. For this purpose, we proceed to locate, collect and analyze the factors which either on short-term or on long-term may have an impact on the shaping of this variable. By achieving that we are able to model the Pension Expenditures and make forecasts. The analysis focuses on 20 European countries for which a large amount of data are available including a set of 20 possible explanatory variables for the period 2001-2015.

Keywords Pension Expenditures; Modelling and Forecasting; Generalized Linear Models; Principal Component Analysis

Kimon Ntotsis
University of the Aegean, Samos, Greece, e-mail: kntotsis@gmail.com

Marianna Papamichail
National Actuarial Authority of Greece, e-mail: marpamich@yahoo.gr

Peter Hatzopoulos
University of the Aegean, Samos, Greece, e-mail: xatzopoulos@aegean.gr

Alex Karagrigoriou
University of the Aegean, Samos, Greece, e-mail: alex.karagrigoriou@aegean.gr

Estimation of classification rules from partially classified data

Geoffrey McLachlan

Abstract We consider the situation where the observed sample contains some observations whose class of origin is known (that is, they are classified with respect to the g underlying classes of interest), and where the remaining observations in the sample are unclassified (that is, their class labels are unknown). For class-conditional distributions taken to be known up to a vector of unknown parameters, the aim is to estimate the Bayes' rule of allocation for the allocation of subsequent unclassified observations. Whatever the reason for wishing to carry out the estimation on the basis of both classified and unclassified data, it can be undertaken in a straightforward manner by fitting a g-component mixture model by maximum likelihood via the EM algorithm in the situation where the observed data can be assumed to be an observed random sample from the adopted mixture distribution. This assumption applies if the missing-data mechanism is ignorable in the terminology introduced by Professors Little and Rubin. An initial likelihood approach was to use the so-called classification maximum likelihood approach whereby the missing labels are taken to be parameters to be estimated along with the parameters of the class-conditional distributions. However, as it can lead to inconsistent estimates, the focus of attention switched to the mixture maximum likelihood approach after the appearance of the EM algorithm. Particular attention is given here to the results of Professor Brad Efron on the asymptotic relative efficiency (ARE) of logistic regression and their basis for the result of Professor Terry O'Neill on the ARE of the mixture maximum likelihood-based rule estimated from partially classified data. Lastly, we consider briefly some current results with Daniel Ahfock in situations where the missing label pattern is non-ignorable for the purposes of maximum likelihood estimation for the mixture model.

Keywords Bayes' rule of allocation; partially-classified data; mixture models; semi-supervised learning

Geoff McLachlan
University of Queensland, Australia, e-mail: g.mclachlan@uq.edu.au

Classification with imperfect training labels

Timothy Cannings, Yingying Fan, and Richard Samworth

Abstract We study the effect of imperfect training data labels on the performance of classification methods. In a general setting, where the probability that an observation in the training dataset is mislabelled may depend on both the feature vector and the true label, we bound the excess risk of an arbitrary classifier trained with imperfect labels in terms of its excess risk for predicting a noisy label. This reveals conditions under which a classifier trained with imperfect labels remains consistent for classifying uncorrupted test data points. Furthermore, under stronger conditions, we derive detailed asymptotic properties for the popular k-nearest neighbour ($k$nn), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. One consequence of these results is that the $k$nn and SVM classifiers are robust to imperfect training labels, in the sense that the rate of convergence of the excess risks of these classifiers remains unchanged. On the other hand, the LDA classifier is shown to be typically inconsistent in the presence of label noise unless the prior probabilities of each class are equal.

Keywords Label noise; Linear discriminant analysis; Misclassification error; Nearest neighbours; Statistical learning; Support vector machines

Timothy Cannings
University of Edinburg, United Kingdom, email: timothy.cannings@ed.ac.uk

Yingying Fan
University of Southern California, United States, email: fanyingy@marshal.usc.edu

Richard Samworth
University of Cambridge, United Kingdom, email: r.samworth@statslab.cam.ac.uk

Classification with unknown class conditional label noise on non-compact feature spaces

Henry W J Reeve, and Ata Kaban

Abstract We investigate the problem of classification in the presence of unknown class conditional label noise in which the labels observed by the learner have been corrupted with some unknown class dependent probability. In order to obtain finite sample rates, previous approaches to classification with unknown class conditional label noise have required that the regression function attains its extrema. We shall consider this problem in the setting of non-compact metric spaces, where the regression function need not attain its extrema.

In this setting we determine the minimax optimal learning rates (up to logarithmic factors). The rate displays interesting threshold behaviour: When the regression function approaches its extrema at a sufficient rate, the optimal learning rates are of the same order as those obtained in the label-noise free setting. If the regression function approaches its extrema more gradually then classification performance necessarily degrades. In addition, we present an algorithm which attains these rates without prior knowledge of either the distributional parameters or the local density. This identifies for the first time a scenario in which finite sample rates are achievable in the label noise setting, but they differ from the optimal rates without label noise.

This work is due to be published by the Proceedings of Machine Learning Research, Volume 99, Conference on Learning Theory 2019.

Keywords classification; label noise; non-compact metric spaces; minimax rates

Henry Reeve
University of Birmingham, UK, e-mail: henrywjreeve@gmail.com

Ata Kaban
University of Birmingham, UK, e-mail: a,kaban@cs,bham.ac.uk

Supervised classification of long or unbalanced datasets

Laura Anderlucci, Roberta Falcone, and Angela Montanari

Abstract Matrix sketching is a recently developed data compression technique. An input matrix A is efficiently approximated with a smaller matrix B, so that B preserves most of the properties of A up to some guaranteed approximation ratio. In so doing numerical operations on big data sets become faster. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. The statistical properties of sketched regression algorithms have been widely studied previously. We study the performances of sketching algorithms in the supervised classification context, both in terms of misclassification rate and of boundary approximation, as the degree of sketching increases. We also address, through sketching, the issue of unbalanced classes, which hampers most of the common classification methods.

Keywords sketching; supervised classification; unbalanced classes

Laura Anderlucci
University of Bologna, laura.anderlucci@unibo.it

Roberta Falcone
University of Bologna, roberta.falcone3@unibo.it

Angela Montanari
University of Bologna, angela.montanari@unibo.it

Kernel change point detection on the running statistics: A flexible, comprehensive and user-friendly tool

Eva Ceulemans, Jedelyn Cabrieto, Kristof Meers, Janne Adolf, Peter Kuppens, and Francis Tuerlinckx

Abstract In many scientific disciplines, studies have demonstrated that on top of the mean, signaling other types of changes is crucial to better capture and understand an event. For example, in emotion psychology, it has been uncovered that it is not only response patterning (i.e., simultaneous change in means) but also response synchronization (i.e., change in the correlations) that characterize response concordance during an emotional episode. In psychopathology research, recent evidence revealed that changes in three statistics, namely, the variance, autocorrelation and correlation, can serve as early warning signs before relapse to depression. In this presentation, we will present KCP-RS, a change point detection tool that can be tailored to capture changes not only in the means but in any statistic that is relevant to the researcher. KCP-RS implements KCP (Kernel Change Point) detection on the running statistics, a derived time series reflecting the statistics of interest. These running statistics are extracted by sliding a window across the time series, and in each window, computing the statistics value. Next, we will put forward a KCP-RS workflow to guide researchers in how to carry out the analysis when multiple running statistics need to be tracked. Finally, using stocks return data and physiological time series, we will introduce the R package we recently built to make KCP-RS freely and easily accessible to applied researchers.

Keywords change point detection; non-parametric; time series analysis; running statistics; early warning signs

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@kuleuven.be

Jedelyn Cabrieto
KU Leuven, Belgium, e-mail: jed.cabrieto@kuleuven.be

Kristof Meers
KU Leuven, Belgium, e-mail: kristof.meers@kuleuven.be

Janne Adolf
KU Leuven, Belgium, e-mail: janne.adolf@kuleuven.be

Peter Kuppens
KU Leuven, Belgium, e-mail: peter.kuppens@kuleuven.be

Francis Tuerlinckx
KU Leuven, Belgium, e-mail: francis.tuerlinckx@kuleuven.be

School motivation profiles of students in secondary education

Matthijs J. Warrens, and Denise M. Blom

Abstract Because school motivation is positively associated with academic performance, profiles of school motivation dimensions may help explain why some students learn and thrive in school contexts, while others struggle academically. Additional research on school motivation may help design interventions for students that struggle academically. For 13,933 9th grade students from the Netherlands, profiles of school motivation were explored using clustering methods. The aim was to identify school motivation profiles in a four dimensional motivation space, including mastery, performance, social and extrinsic motivation.

There is no consensus in the literature on what clustering methods are best suited for exploring school motivation profiles, and what validity indices are best suited for determining the number of profiles (clusters). The performance of three different clustering methods (k-means, k-medoids, latent cluster analysis) and multiple validity indices (Dunn index, CH-criterion, BIC, AIC, among others) was compared. Instead of focusing on a relatively small numbers of clusters, which is common in clustering applications, all clustering solutions up to 100 clusters were summarized and compared.

Keywords benchmarking clustering methods; school motivation; secondary education

Matthijs J. Warrens
University of Groningen, The Netherlands, e-mail: m.j.warrens@rug.nl

Denise M. Blom
University of Groningen, The Netherlands, e-mail: d.m.blom@student.rug.nl

Probing the nature of psychological constructs with Taxometrics and Latent Class Analysis: The case of children's mental models

Dimitrios Stamovlasis, Julie Vaiopoulou, and George Papageorgiou

Abstract A crucial issue in psychological and educational research is to reveal the nature or the type of a latent variable, that is, to attest if it is categorical or continuous. This challenge is actually a classification problem, where scientists have developed suitable psychometric models, based on different assumptions, and attempt to measure their goodness-of-fit with empirical data. Latent Class Analysis (LCA) is an eminent psychometric method implemented to achieve a typology based on a set of observed variables. A satisfying fit in LCA, however, does not assurance the categorical nature (taxon) of the latent construct. A specialized approach designed to detect the type of the latent structure is Taxometric Analysis (TA). The present paper explicates both methods and presents some empirical applications with children's naïve knowledge on the physical reality, providing also answer for the fundamental hypothesis under study. It is imperative to stress that whether children's mental representations is considered as categorical (taxons) or dimensional is of paramount importance in psychology of leaning, because this assumption actually determines their definitions, their measurement and mainly their pedagogical content knowledge.

Keywords latent class analysis; children's mental models; taxon; taxometrics

Dimitrios Stamovlasis
Aristotle University of Thessaloniki, Greece, e-mail: stadi@edlit.auth.gr

Julie Vaiopoulou
Democritus University of Thrace, Alexadroupolis, Greece, e-mail: jvaiopoulou@gmail.com

George Papageorgiou
Democritus University of Thrace, Alexadroupolis, Greece, e-mail: gpapageo@eled.duth.gr

On the use and reporting of cluster analysis in educational research: A systematic review

Hanneke van der Hoef, Matthijs J. Warrens, and Marieke E. Timmerman

Abstract In applied cluster analysis studies, there is little unification and standardization in both the use of methods and reporting of results. Inadequate or incomplete use and reporting practices may impede replicability, a critical appraisal of findings, and a comparison of results between studies. A first step towards unification and standardization is to assess the current use and reporting practices and summarize these in an overview. Because cluster analysis is applied in a wide variety of fields with notably different goals, such an overview should be formed for each domain separately. In our work, the focus is on educational research. In educational research there is an emerging need to identify academic ability profiles with the ultimate goal to improve appropriate school selection and tailored curricula. As a result of the rising emphasis to study students' academic ability in a multivariate, person-centered context, cluster analytic techniques are becoming increasingly popular in educational research. In the current work, we conducted an extensive search in four electronic databases covering the past decade (2008-2018). This search identified more than 2000 records. After a three phase screening procedure, we included 146 papers in the review. A specific area of focus in this work is the selection of the number of clusters. Which range is being considered? Are plots and statistics provided for all models that were considered or (only) for the final selected model? Which statistics are used to select the number of clusters? Do these statistics tend to agree about the 'best' number of clusters? Is interpretation added as a guidance to select the number of clusters, and if so, how? The current study aims to raise awareness of potential inadequate or incomplete reporting practices. Herewith, we aim to contribute to proper use and reporting of cluster analysis.

Keywords information criteria; latent classes; model fit; model selection; multivariate analysis

Hanneke van der Hoef
University of Groningen, the Netherlands, e-mail: h.van.der.hoef@rug.nl

Matthijs J. Warrens
University of Groningen, the Netherlands, e-mail: m.j.warrens@rug.nl

Marieke E. Timmerman
University of Groningen, the Netherlands, e-mail: m.e.timmerman@rug.nl

Predictive ensemble methods for event time data

Berthold Lausen

Abstract The modelling of the relationship of some response to variables measured on different scales is of interest in many applications. The identification and modelling of interactions or subgroups is a specific challenge. Classification and regression trees (CART) is a popular approach. We review a proposal to provide an unbiased variable selection. Event time data can be analysed by using the Logrank statistic as split criterion and by using a Kaplan-Meier estimator as predicted survival (event time) function at each leaf of the tree.

Trees are known to be weak classifiers. Ensemble of trees, e.g. bagging and random forests, are introduced to derive strong classifiers based on trees. Majority voting, estimated probabilities, average predictions are used to define the prediction of a tree ensemble of classifiers or regression models. A weighted Kaplan-Meier estimator is used as predictor for bagged survival trees. Recently, this proposal was generalised by predicting the distribution for a new observation by aggregating the observations of all leafs of the tree ensemble the new observation belongs to. Being interested in the mean (regression or probability tree) it gives the same prediction rule as before, but it is possible to estimate other characteristics of the distribution as a quantile for example.

Selected tree ensembles can be applied to reduce the size and to improve random survival forests.

Keywords Survival data, Kaplan-Meier estimator, Logrank statistic, selected survival tree ensembles

Berthold Lausen
University of Essex, UK, email: blausen@essex.ac.uk

A cellwise trimming approach to Cluster Analysis

Luis Ángel García-Escudero, Diego Rivera-García, Joaquín Ortega, and Agustín Mayo-Iscar

Abstract It is well known than even a small fraction of outlying measurements can detrimentally affect standard clustering procedures. Trimming outlying observations is a sensible and simple way to achieve robustness in Cluster Analysis and some procedures that allow to trim complete observations or cases are available in the literature. However, trimming complete observations, instead of just trimming the most outlying cells within those observations, can be too extreme in sacrificing a lot of valuable information. This is specially the case when dealing with high-dimensional data in which you cannot expect many observations completely free of outlying cells, even with a very small fraction of outlying cells scattered in the data. To overcome this problem, a cellwise trimming approach based on affine subspace approximations and robust regression techniques is presented. The methodology is particularized to functional clustering where only outlying parts are trimmed within curves. The methodology will be illustrated with simulated and real data sets.

Keywords robust clustering; trimming; cellwise outliers

Luis Ángel García-Escudero
Universidad de Valladolid, España, e-mail: lagarcia@eio.uva.es

Diego Rivera-García
Centro de Investigación en Matemáticas CIMAT, México, e-mail: driver@cimat.mx

Joaquín Ortega
Centro de Investigación en Matemáticas CIMAT, México, e-mail: jortega@cimat.mx

Agustín Mayo-Iscar
Universidad de Valladolid, España, e-mail: agustinm@eio.uva.es

Redundancy analysis for categorical data based on logistic regressions

Jose Luis Vicente-Villardón and Laura Vicente-Gonzalez

Abstract Redundancy analysis (RDA) is one of the many possible methods to extract and summarize the variation in a set of response variables that can be explained by a set of explanatory variables. The main idea is to use multivariate linear regression to explain the responses as a linear functions of the explanatory variables and then use Principal Component Analysis (PCA) or Biplots to visualize the results. When response variables are categorical (binary, nominal or ordinal), classical linear techniques are not adequate. Some alternatives as Distance Based RDA have been proposed in the literature. In this paper we propose versions of RDA based on generalized linear models with logistic responses. The natural visualization methods for the visualization of the proposed techniques are the *Logistic Biplots,* recently proposed. The methods are illustrated with an application to real data.

Keywords categorical data; redundancy analysis

Jose L. Vicente-Villardon
Universidad de Salamanca, Spain, e-mail: villardon@usal.es

Laura Vicente-Gonzalez
Universidad de Salamanca, Spain, e-mail: laura20vg@usal.es

A log-ratio approach to cluster analysis of count data when the total is irrelevant

Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, and Javier Palarea-Albaladejo

Abstract Compositional data are strictly positive multivariate observations carrying relative information. Analysis based on log-ratios between the constituting parts has become the mainstream approach for this type of data. When the total is irrelevant, multivariate count data can be understood as compositional data. However, there are some challenges that make them not fully compatible with ordinary compositional analysis. In particular, the presence of count zeros prevents from applying the log-ratio approach directly.

Some strategies have been considered to analyse multivariate count data using log-ratio methods in this context. One consists of firstly replacing zeros by a sensible small positive amount and then apply the methods as ordinarily. Another approach consists of modelling the data as obtained by compounding a model for the underlying vector of probabilities and a model for the counting process, typically a multinomial distribution. In this second approach, once the distribution has been estimated, the latent variables of this process (the strictly positive probabilities) are modelled using the log-ratio approach. The two most common prior distributions used for count data are the Dirichlet and the log-ratio-normal distributions, resulting in the Dirichlet-multinomial and log-ratio-normal-multinomial compounded distributions respectively. Even though the latter represents a more general model, the estimation of its parameters is currently complex and time consuming.

In this work we propose a procedure for cluster analysis of multivariate count data with zeros using the log-ratio-normal-multinomial distribution. In the estimation process we make use of the Dirichlet-multinomial distribution. The proposed algorithm starts with initial estimates given by the Dirichlet-multinomial distribution which are used to determine the compounded distribution. Then, based on the corresponding posterior distribution, a resampling scheme is applied to analyse properties of final clusters defined on the space of the latent probabilities.

Keywords compound distributions; multivariate count data; log-ratio approach; compositional data

Marc Comas-Cufí
University of Girona, Spain, email: mcomas@imae.udg.edu

Josep Antoni Martín-Fernández
University of Girona, Spain, email: jamf@imae.udg.edu

Glòria Mateu-Figueras
University of Girona, Spain, email: gloria@imae.udg.edu

Javier Palarea-Albaladejo
BioSS, UK, email: javier.palarea@bioss.ac.uk

Doing research and teaching data analysis in Greek higher education

Vicky Bouranta, and Iannis Papadimitriou

Abstract This session will focus on the Past, the Present and the Future of Data Analysis in Greece. The aim of this session is to summarize and highlight the effort done during the last 30 years, in the field of Statistics and other scientific fields, dealing with the application, the development and the promotion of the Data Analysis methods. In addition, summaries of recent research experience and developments will be provided. The emphasis will be placed on the promises and the practical implications of the Data Analysis methods in the analysis and interpretation of everyday life multidimensional data, irrespectively of the scientific field from which they come from. A final touch will be on sharing experience and some ideas on new ways for promoting further the use and the (better) understanding of Data Analysis methods and to strengthen the statistical literacy in general. This plenary session will cover three topics. The first topic is: "Doing Research and Teaching Data Analysis in Greek Higher Education" (based on a survey among the members of GSDA) Iannis Papadimitriou (Emeritus Professor) angelikip1@gmail.com Vicky Bouranta (PhD candidate) vickybouranta@gmail.com The second topic is: "Data Analysis Bulletin" (a literature review), Dimitrios Karapistolis (Emeritus Professor) karap@mkt.teithe.gr Marina. Sotiroglou (PhD candidate) misotiro@polsci.auth.gr The third topic is: "the Past, the Presence and the Future" (round table discussion), Ilias Athanasiadis (Emeritus Professor), athanasiadis@rhodes.aegean.gr Giannoula Florou (Professor) gflorou@teikav.edu.gr Georgia Panagiotidou (Teaching Assistant) vzgp@hotmail.com

Keywords geometric data analysis; Greek Society of Data Analysis; history of statistics

Vicky Bouranta
Aristotle University Thessaloniki,Greece, email: vickybouranta@gmail.com

Iannis Papadimitriou
University of Macedonia, Greece, email: angelikip1@gmail.com

Data Analysis Bulletin: (a literature review)

Marina Sotiroglou, and Dimitris Karapistolis

Abstract This session will focus on the Past, the Present and the Future of Data Analysis in Greece. The aim of this session is to summarize and highlight the effort done during the last 30 years, in the field of Statistics and other scientific fields, dealing with the application, the development and the promotion of the Data Analysis methods. In addition, summaries of recent research experience and developments will be provided. The emphasis will be placed on the promises and the practical implications of the Data Analysis methods in the analysis and interpretation of everyday life multidimensional data, irrespectively of the scientific field from which they come from. A final touch will be on sharing experience and some ideas on new ways for promoting further the use and the (better) understanding of Data Analysis methods and to strengthen the statistical literacy in general. This plenary session will cover three topics. The first topic is: "Doing Research and Teaching Data Analysis in Greek Higher Education" (based on a survey among the members of GSDA) Iannis Papadimitriou (Emeritus Professor) angelikip1@gmail.com V. Bouranta (PhD candidate) vickybouranta@gmail.com The second topic is: "Data Analysis Bulletin" (a literature review), Dimitrios Karapistolis (Emeritus Professor) karap@mkt.teithe.gr M. Sotiroglou (PhD candidate) misotiro@polsci.auth.gr The third topic is: "the Past, the Presence and the Future" (round table discussion), I. Athanasiadis (Emeritus Professor), athanasiadis@rhodes.aegean.gr G. Florou (Professor) gflorou@teikav.edu.gr G. Panagiotidou (Teaching Assistant) vzgp@hotmail.com

Keywords geometric data analysis; Greek Society of Data Analysis; history of statistics

Marina Sotiroglou
Aristotle University Thessaloniki, Greece, email: misotiro@polsci.auth.gr

Dimitris Karapistolis
ATEI, Greece, email: karap@mkt.teithe.gr

The Past, the Presence and the Future of Data Analysis

Ilias Athanasiadis, Giannoula Florou, and Georgia Panagiotidou

Abstract This session will focus on the Past, the Present and the Future of Data Analysis in Greece. The aim of this session is to summarize and highlight the effort done during the last 30 years, in the field of Statistics and other scientific fields, dealing with the application, the development and the promotion of the Data Analysis methods. In addition, summaries of recent research experience and developments will be provided. The emphasis will be placed on the promises and the practical implications of the Data Analysis methods in the analysis and interpretation of everyday life multidimensional data, irrespectively of the scientific field from which they come from. A final touch will be on sharing experience and some ideas on new ways for promoting further the use and the (better) understanding of Data Analysis methods and to strengthen the statistical literacy in general. The first topic is: "Doing Research and Teaching Data Analysis in Greek Higher Education" (based on a survey among the members of GSDA) Iannis Papadimitriou (Emeritus Professor) angelikip1@gmail.com Vicky. Bouranta (PhD candidate) vickybouranta@gmail.com Dimitrios Karapistolis (Emeritus Professor), email: karap@mkt.teithe.gr M. Sotiroglou (PhD candidate), email: misotiro@polsci.auth.gr

Keywords: geometric data analysis; Greek Society of Data Analysis; history of statistics

Ilias Athanasiadis
Aegean University, Greece, email: athanasiadis@rhodes.aegean.gr

Giannoula Florou
Eastern Macedonia and Thrace Institute of Technology, email: gflorou@teikav.edu.gr

Georgia Panagiotidou
Aristotle University Thessaloniki, Greece, email: vzgp@hotmail.com

Clustering and classification of interval time series

Paula Brito, Ann Maharaj, and Paulo Teles

Abstract Interval time series (ITS) occur when real intervals of some variable of interest are recorded as an ordered sequence along time. ITS arise, e.g., when we record minimum and maximum temperature values along time, or the daily range of sea levels in different locations, or low and high values of asset prices in consecutive sessions. Clustering and classification of cross-sectional interval data, as well as of standard (i.e., real-valued) time series have been extensively studied; however, techniques for the analysis of interval time series have not received much attention in the literature. In this work, we propose methods for the clustering and classification of ITS, following different approaches. Using time domain features or wavelet features of the radius and centers series allows representing the ITS in a $n \times p$ data array, to which clustering (hierarchical, k-means, etc.) or supervised classification methods (LDA, QDA,…) may then be directly applied. Following a different approach, the ITS may be compared using adequate distances, and then methods which rely solely on a distance matrix can be used such as hierarchical clustering or KNN classification. To this purpose, point-to-point distances, which compare intervals observed at corresponding time-points, or distances based on interval auto-correlation or on the auto-correlation matrices (that gather the autocorrelation and cross-correlation functions of the ITS upper and lower bounds) may be used. Furthermore, the feature representation of ITS allows for outlier identification, using robust estimation procedures developed for multivariate cross-sectional data. The different alternative approaches are explored and their performances compared for ITS simulated under different set-ups, and illustrated with sea-level and ECG data.

Keywords distance measures; interval time series; interval auto-correlation

Paula Brito
FEP, University of Porto & LIAAD INESC-TEC, Portugal, e-mail: mpbrito@fep.up.pt

Ann Maharaj
Monash University, Australia, e-mail: ann.maharaj@monash.edu

Paulo Teles
FEP, University of Porto & LIAAD INESC-TEC, Portugal, e-mail: pteles@fep.up.pt

Multiple-valued symbolic data clustering using regression mixtures of Dirichlet distributions

José G. Dias

Abstract Symbolic data analysis (SDA) has been developed as an extension of the data analysis to handle more complex data structures. In this general framework the pair observation/variable is characterized by more than one value: from two (e.g., interval-value data defined by minimum and maximum values) to multiple-valued variables (e.g., frequencies or proportions). This research discusses the clustering of multiple-valued symbolic data using Dirichlet distributions. This new family of models explores the parameterization of compositional data in the regression setting, for instance regression/expert models. Results are illustrated with synthetic and demographic (population pyramids) data.

Keywords multiple-valued symbolic data; clustering; mixture models; Dirichlet distribution; regression

José G. Dias
Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Portugal, e-mail: jose.dias@iscte-iul.pt

Visualization of heterogeneity in exploratory meta-analysis

Masahiro Mizuta

Abstract Meta-analysis is a method to integrate multiple research results and is widely used in medicine and sociology. It is evaluated as the highest evidence in medicine. A kind of meta-analysis, exploratory meta-analysis is currently attracting attention, and its purpose is not only to derive statistical evidence but also to search for unknown findings. It is well known that it is inappropriate to simply summarize research results. In meta-analysis, research heterogeneity is a key point. There are Q and I^2 as measures of heterogeneity. These correspond to the distribution of effect sizes. For example, I^2 may be interpreted as low, moderate, and high for 25%, 50%, and 75%. We should decide models to use with these measures as a reference. If heterogeneity exists, the random effect model may be used. However, there are many types of heterogeneity that affect meta-analysis, e.g. the existence of nonlinear relationships, outliers, or clusters. Visualization of research results is important to explore them in an exploratory way or exploratory meta-analysis.

Symbolic data analysis (SDA) was proposed by Dr. Diday in the 1980s and is closely related to meta-analysis. The unit of analysis is SDA is assumed to a class or concept and we can analyze the data taking into consideration their internal variations. Symbolic data analysis can be applied with research results in meta-analysis as concepts. For example, cluster analysis and multidimensional scaling can be applied to similarities by defining the degree of similarity between research results. That means that visualization of research results can be realized. We will consider a method to illustrate the relationship between research results in meta-analysis.

Keywords random-effects model; symbolic data analysis

Masahiro Mizuta
Hokkaido University, Japan, e-mail: mizuta@iic.hokudai.ac.jp

QVisVis: Framework and R toolkit for Exploring, Evaluating, and Comparing Visualizations

Ulas Akkucuk, and Stephen L. France

Abstract The conceptual framework described in this paper, along with associated software, called QVisVis, is designed to help evaluate the performance of different dimensionality reduction techniques. Rather than give a simple overall metric, the framework includes tools to show how well visualizations recover item neighborhoods across a range of neighborhood sizes and in different areas of a visualization graph. The toolkit includes scatter plots, heat maps, loess smoothing, and performance lift diagrams. The overall rationale is to help researchers compare dimensionality reduction techniques and use visual insights to help select and improve techniques. As the model selection task is becoming more important in data mining, the QVisVis tool will provide a practical way to compare a wide selection of available data reduction techniques.

Keywords Dimensionality reduction; mapping; solution quality; model selection

Ulas Akkucuk
Bogazici University, Turkey, e-mail: ulas.akkucuk@boun.edu.tr

Stephen L. France
Mississippi State University, USA, e-mail: sfran@business.msstate.edu

Visual exploration for feature extraction and feature engineering

Adalbert F.X. Wilhelm

Abstract Human movement recognition is an important classification task for both robotics research as well as human disease treatment such as knee rehabilitation. Typical datasets comprise multiple sensor measurements for various participants for different physical activities such as sit, walk, run, but also movement changes, e.g. sit-to-stand, lateral-shuffle-left. Visual exploration of the sensor data allows to extract features that are particular helpful for differentiating between the physical activities. Moreover, proper visual explanation enables the analysist to better understand the overall shape of sensor data and the communalities between different individuals. In this talk, a comparative evaluation of various visualizations is given with respect to their efficiency in extracting features for this purpose. We illustrate this with a data set comprising 22 different human physical activities.

Keywords classification; graphics; data visualization; human activity recognition

Adalbert F. X. Wilhelm
Jacobs University Bremen, Germany, e-mail: a.wilhelm@jacobs-university.de

Multivariable analysis on the use of social media & web 2.0/3.0. Modeling & clustering of users

Evangelia Nikolaou Markaki, and Theodore Chadjipantelis

Abstract Using data reduction method, multiple correspondence analysis in two steps and conjoint analysis we create some clusters of voters profiles and we investigate the use of social media and Web 2.0 & 3.0. This study presents how we manage a big number of variables creating models as well as types of social media users via multivariable analysis. Conjoint Analysis represents a hybrid type of technique to examine dependent relations and combines methods such as Regression or Anova permitting researchers to depict a person's preference about a concept, an idea or a product taking into account different characteristics or factors using an experimental design. The relative importance of each characteristic - factor shows its contribution to the "total preference". Data analysis was also based on Multiple Correspondence Analysis (MCA) in two steps. The index-variables were jointly analysed via Multiple Correspondence Analysis on the so-called Burt table. The main MCA output is a set of orthogonal axes or dimensions that summarize the associations between variable categories into a space of lower dimensionality, with the least possible loss of the original information contained in the Burt table. HCA is then applied on the coordinates of variable categories on the factorial axes. This is now a clustering of the variables, instead of the subjects. The groups of variable categories can reveal abstract discourses. To determine the number of clusters, we use the empirical criterion of the change in the ratio of between-cluster inertia to total inertia, when moving from a partition with r clusters to a partition with r-1 clusters. The new media offer a modern environment of interaction, participation and communication is social life, where the information management and diffusion is open. The study shows that the influence exerted in social media is more related to personal interests and life than social issues and social involvement. So, how social media exert influence in the case of political networking, political marketing and involvement in politics? The results as well as the methodology used can be widely used for strategic political marketing and for professionals that manage the political agenda properly.

Keywords conjoint analysis; multiple correspondence analysis; factor analysis; data reduction; social media

Evangelia Nikolaou Markaki
Aristotle University of Thessaloniki, Greece, e-mail: markakie@polsci.auth.gr

Theodore Chadjipantelis
Aristotle University of Thessaloniki, Greece, e-mail: chadji@polsci.auth.gr

Probabilistic collaborative representation learning

Aghiles Salah and Hady W. Lauw

Abstract Recommender systems have become standard in online applications to guide users in navigating the sea of options offered to them. A prevalent approach to recommendation is Collaborative Filtering (CF), which relies on user-item interactions, also referred to as preference data, such as ratings to learn a user's preferences, and provide her with a short list of items that matches her affinities. Preference data however tends to be very sparse, making it very challenging to estimate and generalize a CF model accurately. One promising solution to alleviate this problem is to leverage auxiliary data that can supplement the lack of user-item interactions, such as item textual descriptions, item relationships, user social network, etc. In this light we introduce Probabilistic Collaborative Representation Learning (PCRL), a new Bayesian generative model for jointly modeling user preferences and deep item feature extraction from auxiliary data. The underlying intuition behind our formulation is as follows. The CF component guides the representation learning (RL) part to focus on extracting item features that are relevant for predicting user preferences. The RL component in turn encourages the CF part to rely on items' auxiliary information to explain preferences thereby supplementing the lack of user-item interactions.

As a Bayesian model, PCRL comprises both a conjugate and a non-conjugate component. It is this latter part that makes inference and learning with this model very challenging. Relying on the recent advances in approximate inference/learning, we derived an efficient variational algorithm to estimate PCRL from observations. Interestingly, the proposed algorithm makes it possible to leverage the conjugate part of PCRL thereby, reducing the variance of the required gradient estimators.

We further demonstrated that the main intuitions behind PCRL's formulation are reflected both empirically and theoretically. While we focus on the CF task as a concrete application, PCRL's scope goes beyond that of recommendation. As an example, this model can be considered to learn jointly from texts and images, for multimodal word representations or image captioning.

Keywords collaborative filtering; Bayesian representation learning; poisson factorization

Aghiles Salah
Singapore Management University, e-mail: asalah@smu.edu.sg

Hady W. Lauw
Singapore Management University, e-mail: hadywlauw@smu.edu.sg

User profiling for a better search strategy in e-commerce website

Putri Wikie Novianti, and Fatia Kusuma Dewi

Abstract As one of important funnels in an e-commerce website, search engine demands to be constantly refined to improve the quality of the shown-products on search engine result page. Information retrieval plays a significant role to capture the relevance between tons of thousand documents and users' query. Other signals, furthermore, are necessarily incorporated to the search strategy to improve the search ranking and further to improve the users experience. This study focuses on the application of machine learning algorithm to find such signals. RandomForest algorithm was implemented to evaluate the importance of the features of interest, to predict the likelihood of potential buyers to place a product into their shopping cart. Prior to predictive modelling, our first challenge was to tackle the severe class imbalance problem, which reached up to the level of 98% on the majority class. The model was then trained and evaluated by cross-validation procedure, resulting in 73% of AUC (sensitivity: 64.44%; specificity: 70.35%). Although the model didn't give a good first impression, we decided to reformulate further our search strategy, by combining the model with the relevancy score resulted from the information retrieval technique (not covered in this study). We evaluated the difference of products' ranking from the top two thousand keywords between the legacy and the proposed search algorithm by Wilcoxon rank sum test and adjusted for the multiple testing comparison with Benjamini-Hochberg correction procedure. The test yielded in a significant difference of products' ranking across three thousand tests. Next, we did an on-line experiment to evaluate the impact of the new proposed algorithm towards the conversion rate of our business metrics. The proposed search strategy brought 2.15% 5.5% relative increase to our click rate and net gross marginal value, respectively. This study shows a practical implementation of feature engineering, data preprocessing, predictive modeling and statistical techniques for a better search strategy in an online retail business.

Keywords random forest; predictive modeling; cross validation

Putri Wikie Novianti, PhD
PT Bukalapak.com, Indonesia, e-mail: putri.wikie@bukalapak.com

Fatia Kusuma Dewi
PT Bukalapak.com, Indonesia, e-mail: fatia.kusuma@bukalapak.com

Comparison of the sharing economy participants' motivation

Roland Szilágyi and Levente Lengyel

Abstract Our research focused on sharing economy (SE) which had been gained more and more ground in recent years and receiving increased media coverage nowadays. The use of sharing economy spread at around the end of the years of 2000 significantly. Different authors define differently the system, and also they analyses the participants motivation in different aspects. The main purpose of SE is to improve the utilization of unused assets. Most of people think that only the economic factors have impact on the participation, but the social and environmental motivations also can be important for the users. Also there are many other factors that can influence the participants. Our study's aim to analyze why people participate in sharing economy activities in Hungary. Our question, is there significant difference between the Hungarians motivation compared with other territories. We use Structural Equation Modelling technique to determine which are the most important motivation factors. The study employs survey data from Hungarian sharing economy users.

Keywords Structural Equation Modelling; sharing economy; motivation analysis

Roland Szilágyi
University of Miskolc, Hungary, e-mail: roland.szilagyi@uni-miskolc.hu

Levente Lengyel
University of Miskolc, Hungary, e-mail: lengyel.levente@uni-miskolc.hu

Classification of suicidal execution area in Japan by areal statistics of committed suicide

Takafumi Kubota

Abstract In this study, high-risk of suicide execute areas where suicide persons were found were identified by using areal statistics of committed suicides to spatially model and visualize the statistics for small areas in various prefectures. Two types of data were used to calculate the suicide rates: the number of suicides by location and time (obtained from the Ministry of Health, Labour and Welfare) and the day-night population rate of each municipality (obtained from the Statistics Bureau, Ministry of Internal Affairs and Communications). Then, choropleth maps were used to visualize the calculated suicide rate. The boundaries and office locations of each municipality were used to create neighborhood relation of the data, generate weights for all pairs of neighborhoods, and compute the local Moran's I statistics for each municipality. Finally, the relationship between the suicide rates and the spatially lagged suicide rates were plotted and comparison maps were drawn to identify the clusters, which correspond to high-risk areas. These findings provide important evidence that is needed to implement suicide countermeasures in Japan.

Keywords Spatial cluster; local Moran's I statistics; visualization; suicide

Takafumi Kubota
Tama University, Japan, e-mail: kubota@tama.ac.jp

Visualization and provision method of meteorological data for Energy Management System

Yoshiro Yamamoto, Kazuki Konda, Hideaki Takenaka, Ken T. Murata, and Takashi Y. Nakajima

Abstract A solar radiation estimator estimates the cloud physics that influence the solar radiation measured by satellite observations, and we can estimate this by using an algorithm based on radiation transfer theory. These calculations were performed by the ultra high-speed radiation transfer calculation method (EXAM) and use a neural network approach. Temperature, wind direction, and wind speed are obtained by using GPV/GSM which is the output data from a global numerical estimate model. We have provided data for two areas: combined Asia and Oceania area, and Japan area. The Asia and Oceania data were calculated every 10 minutes, and the Japan data were calculated every 2.5 minutes. The calculated data is provided in binary form on the Solar Radiation Consortium, but there are few users because the data is cumbersome. In order to promote data utilization in the CREST EMS research group, we constructed a data interface system. In the interface system, data can be downloaded CSV data by specifying latitude and longitude or city name along with the period. There is also a Web API that provides data in JSON format. In addition, it also provides a function as GIS visualization. We examined the problems in providing data interface service to CREST EMS research group, which was confirmed by Himawari 8 by high resolution in time and space. As an example of utilization of weather data, we will also introduce data provision for solar car races at WSC 2017. This work was supported by CREST, Japan Science and Technology Agency.

Keywords Visualization; Data providing; solar radiation

Yoshiro Yamamoto
Research & Information Center, Tokai University, Japan, e-mail: yama@tokai-u.jp

Kazuki Konda
Graduate School of Science and Technology, Tokai University, Japan, e-mail: k.konda0626@gmail.com

Hideaki Takenaka
JAXA, Japan, e-mail: neurolink.sys@gmail.com

Ken T. Murata
National Institute of Information and Communications Technology, Japan, e-mail: ken.murata@nict.go.jp

Takashi Y. Nakajima
Research & Information Center, Tokai University, Japan, e-mail: nkjm@yoyogi.ycc.u-tokai.ac.jp

Spatial perception for structured and unstructured data in topological data analysis

Yoshitake Kitanishi, Fumio Ishioka, Masaya Iizuka, and Koji Kurihara

Abstract Of the hypotheses and testing cycles, how to create particularly good hypotheses efficiently is the key to scientific progress. In recent years, data and information have been accumulated and become huge. These data and information should be used efficiently for hypothesis creation. However, with conventional methods (e.g. hierarchical cluster analysis), it has become difficult to capture the features of the data spatially and to visualize robustly against data update and increase. So topological data analysis (TDA Mapper) is attracting attention as a new visualization technology. TDA Mapper creates a graph with the same topological structure as the original data and displays the features of the data as an easy-to-understand graph. On the other hand, there is also the problem of target data. It is becoming very difficult to create innovative hypotheses with the traditional approach of gaining knowledge only from quantitative data stored in structured databases. Therefore, to increase the amount of information, qualitative data such as texts and images are also included in the analysis object. Problem-solving approach to "how to quantify qualitative data for feature extraction and feature value calculation" has a wide variety of purposes and calculation methods. Replace this with data issues in the pharmaceutical industry. Classification of drugs and prediction of target variables using chemical descriptors and physical property data, which are quantitative data, have been conventionally performed. But to cause discontinuous changes in pharmaceutical research, we should promote to use various qualitative data. In this paper, we report the results of spatial perception using TDA Mapper on quantitative data of drugs (mainly chemical descriptors and physical data) and qualitative data (mainly text data including drug information). We also report on the impact analysis of pretreatment, the characteristics and efficiency of the analysis approach, and further applications.

Keywords topological data analysis; TDA mapper; spatial perception; huge data; quantitative data; qualitative data

Yoshitake Kitanishi
Okayama University, Japan, e-mail: ykitanishi@s.okayama-u.ac.jp

Fumio Ishioka
Okayama University, Japan, e-mail: fishioka@okayama-u.ac.jp

Masaya Iizuka
Okayama University, Japan, e-mail: iizuka@okayama-u.ac.jp

Koji Kurihara
Okayama University, Japan, e-mail: kurihara@ems.okayama-u.ac.jp

Dimensional reduction clustering with modified outcome method

Kensuke Tanioka and Hiroshi Yadohisa

Abstract In randomized clinical trials, it is very important to estimate the causal treatment effect, from these covariates, between intervened group and control group. However, it is difficult to estimate and interpret the treatment effect since each subject is assigned to only one therapy. For the estimation of the treatment effect, modified outcome method (Tian et al., 2014) is very useful method. The advantage of the modified outcome method is that it is easy to estimate the treatment effect by using the notion of randomization. However, if the number of covariates is larger, it becomes difficult to interpret the meaning of the treatment effect even if the sparse estimation such as lasso is used. To overcome this problem, we proposed dimensional reduction clustering with the notion of the modified outcome method. By using the proposed method, we can easily interpret the treatment effect through both the estimated component loadings of covariates and clustering structures.

Keywords randomization; alternative least squares criterion

Kensuke Tanioka
Wakayama Medical University, Japan, e-mail: kensuke.t0628@gmail.com

Hiroshi Yadohisa
Doshisha University, Japan, e-mail: hyadohis@mail.doshisha.ac.jp

Forecasting transportation demand for the U.S. market

Vasilios Plakandaras, Theophilos Papadimitriou, and Periklis Gogas

Abstract In this paper we forecast air, road and train transportation demand for the U.S. domestic market based on econometric and machine learning methodologies. More specifically, we forecast transportation demand for various horizons up to 18 months ahead, for the period 2000:1– 2015:03, employing, from the domain of machine learning, a Support Vector Regression (SVR) and from econometrics, the Least Absolute Shrinkage and Selection Operator and the Ordinary Least Squares regression. In doing so, we follow the relevant literature and consider the contribution of selected variables as potential regressors in forecasting. Our empirical findings suggest that while all models outperform the Random Walk benchmark, the machine learning applications adhere more closely to the data generating process, producing more accurate out-of-sample forecasts than the classical econometric models. In most cases, we find that the transportation demand is driven by fuel costs, except for road transportation where macroeconomic conditions affect transportation volumes only for specific forecasting horizons. This finding deviates from the existing literature, given the support of previous studies to macroeconomic conditions are driving factors of transportation demand. Our work relates directly to decisions on transport infrastructure improvement, while it can also be used as a forecasting tool in shaping transportation-oriented policies.

JEL Classification: C32, C53, L41

Keywords transportation; transportation demand; forecasting; machine learning; Support Vector Regression; LASSO

Vasilios Plakandaras
Democritus University of Thrace, Greece, email: vplakand@econ.duth.gr

Theophilos Papadimitriou
Democritus University of Thrace, Greece, email: papadimi@econ.duth.gr

Periklis Gogas
Democritus University of Thrace, Greece, email: pgkogkas@econ.duth.gr

Money neutrality, monetary aggregates and machine learning

Emmanouil Sofianos, Theophilos Papadimitriou, and Periklis Gogas

Abstract The issue of whether or not money affects real economic activity (money neutrality) has attracted significant empirical attention over the last five decades. If money is neutral even in the short-run, then monetary policy is ineffective and its role limited. If money matters, it will be able to forecast real economic activity. In this study, we test the traditional simple sum monetary aggregates that are commonly used by central banks all over the world and also the theoretically correct Divisia monetary aggregates proposed by the Barnett Critique, both in three levels of aggregation: M1, M2 and M3. We use them to directionally forecast the Eurocoin index: a monthly index that measures the growth rate of the euro area GDP. The data span from January 2001 to June 2018. The forecasting methodology we employ is Support Vector Machines (SVM) from the area of Machine learning. The empirical results show that: a) the Divisia monetary aggregates outperform the simple sum ones and b) both monetary aggregates are able to directionally forecast the Eurocoin index reaching a highest accuracy of 82.05% providing evidence against money neutrality even in the short run.

Keywords Eurocoin; Simple Sum; Divisia; SVM; Machine Learning; Forecasting; Money Neutrality

Emmanouil Sofianos
Democritus University of Thrace, Greece, e-mail: esofiano@econ.duth.gr

Theophilos Papadimitriou
Democritus University of Thrace, Greece, e-mail: papadimi@econ.duth.gr

Periklis Gogas
Democritus University of Thrace, Greece, e-mail: pgkogkas@econ.duth.gr

Forecasting S&P 500 spikes: an SVM approach

Athanasios-Fotios Athanasiou, Theophilos Papadimitriou, and Periklis Gogas

Abstract In this study, we focus on forecasting long-tail events of the S&P500 stock returns. The S&P500 is widely considered as a bellwether for the overall U.S. economy as it encompasses some of the largest -in terms of capitalization- corporations from both the NYSE and the NASDAQ stock exchanges. A timely and efficient forecast of such extreme changes is of great importance to market participants and policy makers. Such spikes may trigger various large scale selling or buying strategies in large portfolios that may have a significant impact in the specific market and the overall economy. We define as "spikes" the events where we have extreme upward or downward changes of the S&P500 index; in our case, we use the returns that fall outside a two-standard deviations band. Moreover, instead of simply using the unconditional overall standard deviation, we employ a GARCH(p,q) model to derive the conditional standard deviation of the returns. This is a more appropriate measure of immediate risk to the market participants than the overall series' standard deviation. Traditional forecasting models that rely on statistical analysis and econometrics, assume that returns follow some standard underlying distribution. These models usually fail to successfully and efficiently accommodate price spikes especially when it comes to forecasting. Instead, in our study, we use the atheoretical and data-driven Support Vector Machines Methodology from the area of Machine Learning. This forecasting approach does not require any initial assumptions on the distribution of the data but rather exploits patterns that may be inhibited in the initial data space. These patterns may become more apparent and exploitable in the resulting feature space. We use daily observations from 01/01/2009 to 27/04/2017. Our overall optimum forecasting model achieved a 70.69% forecasting accuracy for the spikes and 73.25% for non-spikes.

Keywords forecast; machine learning; support vector machines; spikes; S&P 500; GARCH

Athanasios-Fotios Athanasiou
Democritus University of Thrace, Greece, e-mail: aathan@econ.duth.gr

Theophilos Papadimitriou
Democritus University of Thrace, Greece, e-mail: papadimi@econ.duth.gr

Periklis Gogas
Democritus University of Thrace, Greece, e-mail: pgkogkas@econ.duth.gr

Assessing the resilience of the U.S. banking system

Anna Agrapetidou, Periklis Gogas, and Theophilos Papadimitriou

Abstract After the 2007 financial crisis, bank sector resilience came into focus for the regulatory authorities. This paper investigates whether regulatory changes affected positively or negatively the resilience of the U.S. banking system. To do so, we exploit a machine learning based bank failure and stress-testing method proposed recently by Gogas et al. (2018) that provides state-of-the-art forecasting accuracy. We focus our analysis on the period from 2000 to 2016 and test the forecasting accuracy of the proposed model by considering the set of all 5,818 U.S. banks - both solvent and insolvent. Next, we provide an alternative perspective on the assessment of U.S. banking system resilience before and after the financial crisis using the center of mass of the banking system. We highlight the changes in the margin of safety (distance from default) from one year to the next to examine the evolution of the whole U.S. banking system over the 2000-2016 period. The results show, starting in 2004, a significantly narrowing safety margin for the U.S. banking sector that led to the financial crisis of 2007; from 2008-2016, the safety margin widens rendering U.S. banks more resilient to external shocks. We show evidence that this widening of the safety margin is the result of a less fierce competition as banks become less in total number and more prudent.

Keywords Bank failures; Stress testing; Forecasting

Anna Agrapetidou
Democritus University of Thrace, Greece, e-mail: aagrapet@econ.duth.gr

Periklis Gogas
Democritus University of Thrace, Greece, e-mail: pgkogkas@ierd.duth

Theophilos Papadimitriou
Democritus University of Thrace, Greece, e-mail: papadimi@ierd.duth.gr

Gender quotas and electoral outcomes for women in european parliamentary elections

Rachel Gregory

Abstract Many EU member states have at least one form of gender quotas in place directed at either local or national political offices. Within the last two EU Parliamentary election cycles, fourteen out of twenty-eight member states have integrated gender quotas into EU offices as well. The use of gender quotas within EU states shows an increase in positive election outcomes for women, but the effectiveness of quota systems often relates to the type of quota in place and the rewards or consequences for adhering to quota rules. Currently, no studies exist that examine the relationship between gender quotas and European election outcomes, partly due to the complexity of gender quota systems in place for EU elections that differ from national and local within-state elections.

Using data from the May 2019 EU elections, this research seeks to fill this gap by measuring women's electoral outcomes between countries with gender quotas and those without and between classifications of quotas. First, by examining outcomes for women overall in election to EU parliament, this research will note any gaps between the election of men and women across member states. Then, in order to determine the impact of gender quotas across EU member states, a comparison between states with gender quotas and those without EU-level quota systems analyses differences between quota and non-quota systems.

After creating a framework for understanding discrepancies between gender in the 2019 EU elections, this research primarily focuses on developing a classification method for countries with gender quota rules for EU elections by clustering states based on both formal and informal quota rules. Using a statistical approach to gender quota classification instead of the theoretical approach common in literature on gender quotas, develops a system of quota classification based on practical applications and outcomes in order to determine if a difference in the type of quota system impacts electoral outcomes.

As the European Union assesses policy options in confronting issues of representation within EU institutions, the impact of varying electoral regimes highlights potential approaches to increase women's participation in European bodies. Rather than isolating the impact of women's electoral outcomes by only states with or without gender quotas, understanding differences between types of quota systems isolates policy measures indicative of increased democratic representation.

Keywords gender quotas; clustering; EU elections

Rachel Gregory
University College Cork, Ireland, rachel.gregory@ucc.ie

What was really the case? Party competition in Europe at the occasion of the 2019 European Parliament Elections

Theodore Chadjipadelis, and Eftichia Teperoglou

Abstract The main aim of the paper is to analyze political competition in EU member states at the occasion of the 2019 European Parliament elections. At the core of our analysis are both the priorities of the national parties campaigning for the 2019 European elections and the manifestos of the transnational party groups, each consisting of national member parties from the 28 member states of the European Union. By comparing the major priorities of national actors/parties and those of the European political groups, we will be able to gauge out whether they share different or same dimensions of policy. More broadly, we will depict whether the dynamism in policy competition at the national level affects EP political groups or *vice versa*. The analysis is implemented through the use of correspondence analysis. Through this approach the axes of political competition are realized.

Keywords European elections 2019; European Parliament; policy positions; party cohesion; party competition

Theodore Chadjipadelis
Aristotle University of Thessaloniki, Greece, e-mail: chadji@polsci.auth.gr

Eftichia Teperoglou
Aristotle University of Thessaloniki, Greece, e-mail: efteperoglou@polsci.auth.gr

First-time voter in Greece: Views and attitudes of youth on Europe and democracy

Georgia Panagiotidou, and Theodore Chadjipadelis

Abstract The research analyzes the views, attitudes and values of young people in Greece with the use of multivariate analysis methods: MCA, HCA and conjoint analysis. The objective is to produce a semantic map of the political behavior of young Greek first-time voters and identify important factors which determine their vote.

The survey was conducted during April 2019 with a sample of 4.500 young pupils and students in Greece age 17-25. The research project aims to identifiy the attitudes of young citizens on European Union, evaluate their degree of political knowledge, their ways of political mobilization, their position on the "Right-Left" scale, their intention to vote or abstain in the European elections, their interest in politics, their degree and sources of information, their personal values and their perception of democracy. Furthermore, the analysis focuses on evaluating the significance of the following criteria in vote choice: parties (ideological identity and worldview), persons (candidates and political personnel), issues (policy and program proposals) and cooperation (position and willingness to search for common political governance). Conjoint analysis can detect the true weight of different factors, in this case the criteria of young people based on which they choose to vote. The use of conjoint in political surveys offers an innovative way to approach questions where the respondent is not able to give a direct reply or does not express a strong preference among different factors. The research intends to classify (HCA) subjects using multiple variables of political behavior in order to identify a basic behavioral typology. In addition to this, the analysis proceeds to creating a semantic map of the political behavior of young people (MCA). This semantic map will provide a visualization of the relationships among groups of young voters and various political characteristics and attitudes, interpreting the political behavior of young people today, also known as generation Z.

Keywords hierarchical clustering; multiple correspondence analysis; conjoint analysis; political analysis; political behavior; young voters; European elections; semantic map

Georgia Panagiotidou
Aristotle University of Thessaloniki, Greece, e-mail: gvpanag@polsci.auth.gr

Theodore Chadjipadelis
Aristotle University of Thessaloniki, Greece, e-mail: chadji@polsci.auth.gr

Developing a model for the analysis of the political programmes

Panagiotis Paschalidis and Theodore Chadjipadelis

Abstract This model focuses on political programmes and not political discourse in general. Political programmes present the particularity of being more easily distinguishable, identifiable and quantifiable because of the fact that in most occasions they are developed and presented by political parties or candidates in an autonomous manner (print or digitally) along the other elements that characterize electoral activity (i.e. political identity, news feed, statements, media coverage…).

The key feature of our approach is the methodological analysis of the three-way table that will contain the following three categories of data: a) a general overview of policies along precise and pre-determined number of general thematic areas (i.e. economy, culture, society, administration…), b) the identification of each candidate's or political party's policies along the thematic categories as well as the cross-examination of all candidates' and parties' policies in a combined view and c) the classification of all policies (taken individually or combined) according to their symbolic connotations when it comes to negative/neutral and positive dispositions.

The implementation of this model presupposes the development of a lexicon that will provide key data, clarifications and definitions that will enable the analysis during its various stages (data collection, codification, production of results, table formulation). This lexicon will contain the following data: a) the identification of all lists, b) the identification of a fixed number of thematic categories that will function as matrices for data collection and codification and c) definition of negative, neutral and positive dispositions.

The most challenging part for the production of this lexicon and of the analysis model in general is the definition of the fixed thematic/ policy categories. Only a consistent definition will guarantee a pertinent collection of data and their subsequent codification on the basis of a software for analysis.

The above procedure is applied to Municipal elections in Thessaloniki area. To overcome such difficulties at the municipal level, our approach proposes the identification of the policy categories according to the Municipality's organization structure which is divided into precise areas of responsibility.

The interest of this model will be the possibility for detecting differences and similarities in political programmes in a combined view which pays attention to different policy areas. More generally, it will provide a view on the policy agenda and its priorities in the context of the political programmes.

Keywords political programme; elections; analysis; three-way table; lexicon; thematic categories; policies

Panagiotis Paschalidis
Aristotle University of Thessaloniki, Greece, e-mail: p.panos@lycos.com

Theodore Chadjipadelis
Aristotle University of Thessaloniki, Greece, e-mail: chadji@polsci.auth.gr

How the undecided voters decide?

George Siakas

Abstract How pre-elections polls influence electoral behavior? Actually, there is not an indisputable response on this question. Some researchers do believe that pre-elections polls do not have in reality any effect on voters' decision, while some argue that they indeed influence their behavior. Thus, the later are not in position to appoint a clear pattern on the direction of this influence. One of the most common questions during elections period is when and how the undecided voters will formulate their decision; will they eventually abstain, would they favor the leading party on the polls, or would they sympathize the most unfavorable? In order to explore the undecided voters' electoral behavior, a panel with undecided voters, which members will be selected during two repeated cross-section surveys fielded during the pre-elections period, will be formed. Panel members will be regularly contacted during the elections period but before the elections day -and especially before the exit polls announcement on the day of the upcoming euro-elections (May 26th, before 7pm)- in order to reveal beliefs, attitudes and report their electoral preference. This study will provide some useful insights in the direction of exploring the undecided voters' electoral behavior.

Keywords survey polls; electoral behavior; euro elections; turnout; undecided voters; panel

George Siakas
University of Macedonia Research Institute, Greece, e-mail: siakas@uom.edu.gr

Improving the performance of Japanese authorship attribution with phonetic related information

Hao Sun, and Mingzhe Jin

Abstract Authorship attribution, the science of inferring the author of anonymous documents, has wide applications on text, music, and source code. The typical authorship attribution model consists of two steps, which are stylometric feature extraction and statistical model application. A stylometric feature is a dataset that captures the writing style of the author. These stylometric features are applied to statistical models to determine which author wrote the anonymous text. The proposed stylometric features for European languages were widely divided into four groups, which are the lexical (word n-grams and vocabulary richness) group, the character (character n-grams) group, the syntactic part-of-speech (POS) group, and the semantic (synonyms) group. For Japanese authorship attribution, the position of commas, particles n-grams, and phrase patterns were indicated as useful stylometric features. Nevertheless, there was a long absence of stylometric features from the point of view of phonology. Recent studies have focused on the importance of syllables in authorship attribution. In this study, we further divide syllables into phonemes and discuss where the phonemes n-gram model includes necessary information for Japanese authorship attribution. Our method consists of four steps. Firstly, we created a corpus that contains 400 novels written by 20 Japanese modern novelists. Secondly, we applied the Japanese morphological analyzer called MeCab to separate Japanese sentences into morphemes and converted all the Chinese characters and Hiragana into Katakana. All the 35 phonemes were counted according to a katakana-phoneme comparative table. The phonemes used in this study were /a, i, u, e, o, k, g, sh, s, z, j, t, ch, ts, d, n, h, f, b, p, m, r, w, y, N, q, :, ky, gy, ny, hy, by, py, my, and ry/, where "N" represents a syllabic nasal consonant, ":" represents a long vowel, and "q" represents a double consonant, respectively. Then we built phoneme n-gram models from the counted data. Other stylometric features, such as comma position, POS n-grams, particle n-grams, and phrase patterns were also extracted from the texts for comparison. Finally, we showed the performance of phoneme n-grams on supervised algorithms includes high-dimensional discriminate analysis (HDDA), logistic model tree (LMT), support vector machine (SVM), and random forest (RF). The evaluation criterion for the supervised algorithms is the F-measure, which is the harmonic mean of recall and precision. Our result indicates that phoneme n-grams improve the performance of the authorship attribution model when it is connected to other stylometric features.

Keywords authorship attribution; phoneme; n-gram model

Hao Sun
Graduate School of Culture and Information Science, Doshisha University, Japan, e-mail: hsun@mail.doshisha.ac.jp

Mingzhe Jin
Faculty of Culture and Information Science, Doshisha University, Japan, e-mail: mjin@mail.doshisha.ac.jp

Double helix multi-stage text classification model to enhance chat user experience in e-commerce website

Fiqry Revadiansyah, Abdullah Ghifari, and Rya Meyvriska

Abstract One of the most crucial problems in an e-commerce website is asynchronized product stock from the seller's product page site to their real-world inventory. Having multiple stores for more than one e-commerce may complicate several sellers to update their product stocks automatically. The chat feature is used to bridge the communication between a buyer and a seller, which then be used to detect emptiness product stock from the seller's text messages. Therefore, the natural language understanding algorithm is proposed to figure out particular emptiness product stock intents on the text conversation between the sellers and the buyers. This study emphases on text-based classification machine learning, where chat messages documents are weighted using three archetypes, such as unigram TF-IDF with Chi-squared feature selection, Word2Vec, and Fasttext weights. We proposed many machine learning models algorithms, which run in parallel to map the suitable intent from two sides chat users. As we validated those models using cross-validation method, the best models tuned by bayesian and random search hyperparameter tuning algorithm by AUC selection. Extreme gradient boosting model using Word2Vec weight resulted as the best sellers intent classification model, which yield 84,49% of AUC score (recall: 92,89%, precision: 79,49%). Moreover, the Random Forest model using TF-IDF weight and Chi-squared feature selections resulted as the best buyers intent classification model, which yield 84,92% of AUC score (recall: 92,98%, precision: 79,49%). These best models worked together as a double helix, in order to determine an emptiness product stock intent both from the sellers and the buyers by matching them in line. This study reveals how chat user experience could be improved by machine learning from conversation intent of text-based data, which starting from data retrieval process, feature engineering, data preprocessing, multi-stage text classification and model combination technique, as a breakthrough of an automation process of online product stock management.

Keywords natural language understanding; multi-stage text classification; double helix model

Fiqry Revadiansyah
PT Bukalapak.com, Indonesia, e-mail: fiqry.revadiansyah@bukalapak.com

Abdullah Ghifari
PT Bukalapak.com, Indonesia, e-mail: abdullah.ghifari@bukalapak.com

Rya Meyvriska
PT Bukalapak.com, Indonesia, e-mail: rya.meyvriska@bukalapak.com

Latent dimensions of the museum experience: the role of the online reviews

Melisa L. Diaz, and Anna Calissano

Abstract Social media platforms have been increasingly used by museums to promote and spread activities and knowledge, but the capacity of systematically receiving knowledge back from visitors through these platforms is weak instead. This study uncovers the latent dimension of the TripAdvisor online reviews of the 30 most visited state museums and cultural heritage sites across Italy. For this purpose, a text mining technique, the Latent Dirichlet Allocation (LDA) has been used. LDA is a Bayesian model that calculates the probability of a review to belong to a latent topic, detecting the main subjects of each review. The overall distribution of the latent topics on the reviews of a single museum gives a silhouette of the museum built up simply by the electronic word of mouth. The result shows, that the experience shared on the online reviews is composed by an amount of emotion, remarks on the features of the site, often a set of recommendations for other visitors and observations regarding the context of the place. To delineate better the similarities and differences across the museums and cultural heritage profiles, a k-mean clustering analysis for compositional data was performed. This study is a first step to understand how the cross-fertilization between the museums and their visitors shapes constantly the character of the cultural institutions.

Keywords Museum experience; online reviews; TripAdvisor; latent dirichlet allocation (LDA)

Melisa Lucia Diaz Lema
Politecnico di Milano, Italy, e-mail: melisaluciad.diaz@polimi.it

Anna Calissano
MOX Laboratory for Modeling and Scientific Computing, Politecnico di Milano, e-mail: anna.calissano@polimi.it

A corpus-based approach to explore the stylistic peculiarity of Kouji Uno's postwar works

Xueqin Liu, and Mingzhe Jin

Abstract Literature style usually refers to authors' distinctive writing habits and is the essential element that makes a literary work unique. Stylometry provides reliable statistical methods and data for an objective stylistic analysis, and many stylistic features have been identified and exploited as style markers in author profiling or authorship attribution. This study aims to investigate the stylistic peculiarity of Kouji Uno's postwar works in comparison with contemporary writers by using a quantitative analysis. Kouji Uno is a well-known Japanese littérateur, whose creative activity was interrupted twice due to a mental illness and World War II. Literary critics argued that Uno's style of postwar works is significantly different from other writers. In this study, we create a digital corpus consisting of a total of 149 novels written by Kouji Uno and ten other contemporary writers. After compiling the corpus, a principal component analysis was carried out using three types of stylistic features: the unigram part-of-speech (POS) tagger, the placement of commas, and the length of elements between punctuation marks. We extracted the data of comma placements from two aspects: the morpheme and the POS before the comma. For the calculation of the length of the elements between punctuation marks, we considered three types of length: characters, kana, and morphemes. Finally, using scatter plots we found that Uno's novels were located far from those of other writers and clarified that the most important factor contributing to the distinction between Uno and other writers' literary styles was the usage of punctuation marks. Uno preferred to use punctuation marks, particularly commas, in works published after the war. Although there are no significant differences in comma placements, the length of elements between punctuation marks was observed to be shorter in works by the other writers. Therefore, the peculiarity in Uno's literary style was found to be driven by special usage of punctuation marks, as revealed by the results of the quantitative analysis. In addition, the analysis results revealed that the novels written by Uno have similar usage of punctuation marks to Osamu Dazai and Ton Satomi.

Keywords Kouji Uno; literature style; principal component analysis; punctuation marks

Xueqin Liu
Doshisha University, Japan, e-mail: xliu@mail.doshisha.ac.jp

Mingzhe Jin
Doshisha University, Japan, e-mail: mjin@mail.doshisha.ac.jp

The analyses of the WoS data on network clustering

Anuška Ferligoj, Vladimir Batagelj, and Patrick Doreian

Abstract There is a large literature on network clustering. We collected bibliographic data for the network clustering literature including both community detection and blockmodeling works through to February 22, 2017. The primary data source was the Web of Science. From the obtained data we created a citation network among works. In addition, we included data on authors, journals and keywords to generate some two-mode networks featuring works × authors, works × journals, and works × keywords. The boundary problem is discussed as was a treatment ensuring the studied citation network is acyclic. Lists of the most prominent journals where works in the network clustering literature appeared were created. Components of the studied network were identified and examined. The CPM path through the main component was identified. It revealed a clear transition from the social network part of the literature to the community detection part. The key-route paths revealed the same transition but with more works and a more nuanced view of it. Ten link islands, as clusters, were identified. Detailed discussions were provided for four including one with a clear distinction between the community detection and social networks literatures as being connected through a cut.

Keywords social network analysis; clustering; network clustering; bibliometrics; Web of Science; main path analysis; link islands

Anuška Ferligoj
University of Ljubljana, Slovenia, email: anuska.ferligoj@fdv.uni-lj.si

Vladimir Batagelj
University of Ljubljana, Slovenia, email: vladimir.batagelj@fmf.uni-lj.si

Patrick Doreian
University of Ljubljana, Slovenia, email: pitpat@pitt.edu

Approximate core-and-shell supercluster in statics and dynamics

Boris Mirkin, and Ivan Rodin

Abstract We present a new method for cluster analysis that finds a "composite" supercluster to consist of two non-overlapping parts: its core and shell, at any network (weighted graph) data. The core of a supercluster is a set of elements related to each other much stronger than to its shell. The shell comprises elements of the supercluster that are related to each other too, but not as tight as the core elements. We use an approximation model in which each of the two, the core and the shell, is assigned with its own intensity parameter. More precisely, given a set of entities I and similarity matrix A=(a(i,j)), i,j from I, we approximate A, possibly shifted to an average origin value, with a supercluster matrix B=(b(i,j)). The common entry b(i,j) is defined as b(i,j)=\alpha*c(i)*c(j)+\beta*s(i)*s(j) where c and s are binary indicator vectors so that c(i)=1 if i belongs to the core to be found, and c(i)=0, otherwise; and s(i)=1 if i belongs to all of the supercluster, and 0, otherwise; while \alpha and \beta are positive intensity weights. All four items in b(i,j) are to be determined so that the least squares one-cluster criterion, the sum of squares of differences a(i,j)-b(i,j), is minimized. This criterion allows us to formulate a method at which a suboptimal cluster and its core are found first, after which both are updated according to adaptive intensity values. We extend this approach to network data changing over time. With the temporary network data, we define and find supercluster so that its core does not change over time, while shells of the dynamic supercluster may differ at each time period. We provide a similar approximation model for the concept of composite dynamic supercluster and derive formulas for its parameters. We present and test an iterative algorithm for finding superclusters one-by-one. This approach is applied to the analysis of a dynamic co-citation network comprising 41 researchers who were active and participated in biannual IFCS Conferences in 1997-2016. This network is organized as 10 biannual co-citation graphs. We found five superclusters at the network, both cores and shells.

Keywords cluster; supercluster; temporary network

Boris Mirkin
National Research University Higher School of Economics, Moscow, Russia, e-mail: bmirkin@hse.ru

Ivan Rodin
Skolkovo Institute of Science and Technology, Moscow RF, e-mail: ivvrodin@gmail.com

Trust your data or not - Standard remains Standard (QP); implications for robust clustering in social networks

Immanuel Bomze, Michael Kahr, and Markus Leitner

Abstract We focus on Clustering in Social Networks applications in a Machine Learning context. A fundamental problem arising in social network analysis regards the identification of communities (e.g., work groups, interest groups), which can be modeled with the framework of a so-called Standard Quadratic Optimization Problem (StQP), where a possibly indefinite quadratic form is maximized over the standard probability simplex. However the problem data are uncertain as the strength of social ties can only be roughly estimated based upon observations. Therefore the robust counterpart for these problems refers to uncertainty only in the objective, not in the constraints. It turns out that for the StQP, most of the usual uncertainty sets do not add complexity to the robust counterpart.

Keywords Robust optimization; Quadratic optimization; Graph clustering

Immanuel Bomze
ISOR/VCOR & ds:UniVie, Universität Wien, Austria, e-mail: immanuel.bomze@univie.ac.at

Micbael Kahr
ISOR/VCOR, Universität Wien, Austria, e-mail: m.kahr@univie.ac.at

Markus Leitner
Vrije Universiteit Amsterdam, The Netherlands, e-mail: m.leitner@vu.nl

Classifying users through keystroke dynamics

George Peikos, Ioannis Tsimperidis, and Avi Arampatzis

Abstract In this paper we propose a method for classifying users based on some inherent or acquired characteristics, through the way they type. Billions of users are connected daily on the Internet for reasons of work, communication, entertainment, and education. Many times they remain completely anonymous, concealing or misrepresenting some of their traits, such as gender and age, either through negligence, or because they do not have to declare them, or because they have some fraudulent purpose. Knowledge of certain characteristics of fully unknown users can help firstly to warn unsuspecting users of being cheated, secondly to obtain useful information about the identity of the suspect in cases of cybercrime, and thirdly to facilitate users to exploit Internet services. To accomplish this, we use features of keystroke dynamics, which is the detailed recording of user actions on the keyboard, such as the time that a key is remained pressed and the time elapsed between the using of two consecutive keys. The use of keystroke dynamics features for user classification firstly ensures that personal and/or sensitive user data will not be revealed as the method focuses on how users type, not what they type, secondly ensures that the proposed method will be independent of the typed language, since these features are not related to specific words and phrases of any spoken language, and thirdly ensures that the method will not have particular requirements in hardware and data, since it involves the most simple and frequent form of communication between Internet users, the text, but also the most common communication device, the QWERTY keyboard, physical or virtual. For the purposes of this study, 110 users were recorded during the daily use of their computer and 362 log files were collected, each containing data from approximately 3,500 keystrokes. From these log files, the necessary keystroke dynamics features were extracted, and with the help of 5 well-known machine learning models, users were classified to identify some of their traits. In particular, the age group, the dominant hand and the educational level of a user, between 4, 3 and 5 classes, respectively, were sought and the success rate reached 87.6%, 97.0%, and 84.3%, respectively.

Keywords Keystroke dynamics; user characteristic classification; machine learning; data mining; feature selection; information gain

George Peikos
Democritus University of Thrace, Greece, e-mail: georpeik1@ee.duth.gr

Ioannis Tsimperidis
Democritus University of Thrace, Greece, e-mail: itsimper@ee.duth.gr

Avi Arampatzis
Democritus University of Thrace, Greece, e-mail: avi@ee.duth.gr

K-means, spectral clustering, or DBSCAN: a benchmarking study

Irene Cho, Nivedha Murugesan, and Cristina Tortora

Abstract We perform a benchmarking study to compare three clustering algorithms: K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Spectral Clustering. In an effort to be thorough, we utilize two simulated and three empirical datasets. The first is 'multishapes', a simulated dataset comprising five clusters two of which are not elliptical, which is often used to compare clustering methods. The second one is a two-dimensional simulated dataset that forms a spiral. The remaining three datasets are empirical in nature: (1) a small bike shop dataset including shop names as the variables and the proportions of shop sales as observations; (2) a midsize forest fire dataset containing information about burned areas in the northeast region of Portugal; and (3) a large Airbnb dataset regarding the Airbnb listings located in Santa Clara County, California. The clustering results obtained using the three methods are compared using Adjusted Rand Index (ARI) values as well as silhouette plots. Furthermore, we consider the runtimes of the three clustering techniques for a complete assessment. After performing all three clustering methods on each of the five datasets and comparing the results, we identify the advantages and disadvantages of the algorithms. Although it is difficult to choose one method that performs best on all types of datasets, we find that DBSCAN should generally be reserved for non-convex data with well-separated clusters or for data with many outliers that need to be filtered. However, when it comes to data without any patterns, we believe that K-means or Spectral Clustering might be able to achieve better results.

Keywords Spectral Clustering, K-means, DBSCAN

Irene (Sisang) Cho
San Jose State University, sisang.cho@sjsu.edu

Nivedha Murugesan
San Jose State University, nivedha.murugesan@sjsu.edu

Cristina Tortora
San Jose State University, cristina.tortora@sjsu.edu

Benchmarking minimax linkage

Xiao Hui Tai, and Kayla Frisoli

Abstract Minimax linkage was first introduced in 2004 as an alternative to standard linkage methods used in hierarchical clustering. Minimax linkage relies on distances to a prototype for each cluster; this prototype can be thought of as a representative object in the cluster, hence improving the interpretability of clustering results. In 2011, a subsequent paper analyzed the properties of minimax linkage, popularizing it within the statistics community. The 2011 paper compared minimax linkage to standard linkage methods, making use of five data sets and two evaluation metrics, distance to prototype and misclassification rate (although not all metrics were used on all data sets). In an effort to expand upon this work and evaluate minimax linkage more comprehensively, our benchmark study analyzes additional well-described data sets (empirical and simulated) and evaluates the clustering results for all data sets on all metrics (distance to prototype, misclassification rate, precision and recall). We decided to additionally include the metrics of precision and recall to provide a fairer comparison when class imbalance exists. For full reproducibility, we make all code and data publicly available through an R package on GitHub (xhtai/clusterTruster). We find that minimax linkage often (but not always) produces the smallest distances to prototypes, meaning that minimax linkage produces clusters where objects in a cluster are tightly clustered around their prototype. This is true across a range of values for the total number of clusters (k), although this is not always the case. Special attention should be paid to the case when k is the true known value, as this is arguably the most relevant case in practice. For true k, minimax linkage does not always perform the best in terms of all evaluation metrics studied, including distance to prototype. Some of these results counter the claims of the 2011 paper. Our paper (available on arXiv) was motivated by the IFCS Cluster Benchmarking Task Force's call for neutral clustering benchmark studies and the corresponding white paper, which put forth guidelines and principles for comprehensive benchmarking in clustering. Our work is designed to be a neutral benchmark study of minimax linkage.

Keywords minimax linkage; hierarchical clustering; benchmark analysis

Xiao Hui Tai
Carnegie Mellon University, tai.xiaohui@gmail.com

Kayla Frisoli
Carnegie Mellon University, kfrisoli@stat.cmu.edu

Benchmarking in cluster analysis for mixed-type data

Madhumita Roy, Jarrett Jimeno, and Cristina Tortora

Abstract There are many instances of benchmarking in cluster analysis with continuous data, but only a few with mixed-type data. However, in various real-world applications, variables are often of mixed-type. Therefore, in this paper, we explore the process for benchmarking various clustering methods on high-dimensional simulated mixed-type data sets. Given that many classical clustering methods, such as k-means, are known to work primarily on continuous datasets, we simulate high-dimensional mixed-type data sets, subsequently preprocess each data set by applying multiple correspondence analysis to its categorical part, and finally subject the full preprocessed data set to several clustering methods. Specifically, we use k-means, fuzzy k-means, probabilistic distance clustering, and clustering based on a mixture of multivariate Student's t distributions. In the recent literature, new methods for clustering mixed-type data without preprocessing (such as KAMILA) also begin to emerge. For this purpose, we also apply KAMILA to the raw mixed-type data and compare the results with those of the previously mentioned techniques as applied to the preprocessed data using multiple correspondence analysis. As quality criteria we use the adjusted Rand index and the average number of algorithm iterations.

Keywords mixed-type data; multiple correspondence analysis; KAMILA

Madhumita Roy
San Jose State University, madhumita.roy@sjsu.edu

Jarrett Jimeno
San Jose State University, jarrett.jimeno@sjsu.edu

Cristina Tortora
San Jose State University, cristina.tortora@sjsu.edu

Comparison of dimensionality reduction and cluster analysis methods for high dimensional datasets

Jingfei Gong, Yuwen Luo, and Cristina Tortora

Abstract Clustering is a type of unsupervised learning that groups a set of data points into clusters so that points within each cluster are similar to each other, while points from different clusters are dissimilar. Real life data tend to be high dimensional and clustering them can be very challenging. High dimensional data may contain attributes that are not required for defining clusters but that increase the computational cost. Moreover, data may also contain irrelevant noise dimensions that do not contain information. Dimension reduction is the transformation of high dimensional data into a meaningful representation of reduced dimensionality based on the intrinsic dimensionality of the data. Dimension reduction can be mainly divided into projection and manifold learning. Principal component analysis (PCA) and independent component analysis (ICA) are very popular projection techniques, while the most common manifold learning techniques are Uniform Manifold Approximation and Projection (UMAP) and t-distributed stochastic neighborhood embedding (t-SNE). In cluster analysis, two of the most common approaches are the distance-based one (of which K-means is a frequently used instance), and the model-based one (of which clustering based on the Gaussian Mixture Model (GMM) is a classical instance). In this paper, we apply four-dimension reduction techniques (PCA, ICA, UMAP, and t-SNE) followed by two cluster analysis methods (K-means and GMM) to three sets of high dimensional empirical data and two sets of high dimensional simulated data. In order to compare the performances of the selected combinations of techniques, we choose the Adjusted Rand Index (ARI) and running time as evaluation criteria; in addition, we also visualize the data in the reduced spaces. We find that the UMAP dimensionality reduction technique in conjunction with GMM cluster analysis outperforms the other combinations of techniques in terms of the two evaluation criteria under study. Regarding the visualization, we notice that points within each cluster are more concentrated and less overlapping among clusters in UMAP plots than in plots for the three other dimension reduction methods, indicating that UMAP performs better.

Keywords clustering; dimension reduction; high-dimensional data

Jingfei Gong
San Jose State University, jingfei.gong@sjsu.edu

Yuwen Luo
San Jose State University, yluoscience@gmail.com

Cristina Tortora
San Jose State University, cristina.tortora@sjsu.edu

Evaluation of text clustering methods and their dataspace embeddings: an exploration

Alain Lelu, and Martine Cadot

Abstract The present contribution to the first Neutral Cluster Benchmarking Challenge is limited to evaluating *text* clustering solutions. As linguistic pre-processing is outside our scope, our starting point consists of three varied open-access corpora, which have been reduced to three sets of raw term occurrence vectors by means of the same publicly available software. We investigated the influence of preprocessing the occurrence data by firstly truncating the size of the vocabulary to a few fixed quantiles, and by subsequently transforming the raw dataspaces into several derived dataspaces that are explicitly or implicitly used in the context of clustering methods. We subjected the preprocessed data to several clustering methods, including classic K-means and hierarchical agglomerative methods, as well as more recent spectral, kernel and graph clustering methods, non-negative matrix decomposition, and latent Dirichlet allocation. As our approach is exploratory, we have not run more than 450 combinations of corpus × term truncation × type and dimensionality of derived dataspace × clustering method, far from all possible combinations of the cited elements. Four usual evaluation indices have been used for comparing the resulting cluster structures with man-made "ground-truth" classes: Normalized Mutual Information, the Adjusted Rand Index, the F-score and the Purity score. The results show both a confirmation of well-established combinations, and good performances of unexpected combinations, mostly in spectral and kernel dataspaces. A disappointing observation is that clear winning combinations emerge for each test corpus, but not for all three together, where the overall best (or least bad) combination seems to be spectral K-means in a Correspondence Analysis factor space. The rich materials resulting from all these runs include a wealth of intriguing facts, which need further research on the specifics of a text corpus in relation to clustering methods and derived dataspaces.

Keywords clustering; evaluation; dataspace

Alain Lelu
rtd, Université de Franche-Comté, alelu@orange.fr

Martine Cadot
LORIA, martine.cadot@loria.fr

Clustering transformed compositional data using K-means, with applications in gene expression and bicyle sharing system data

Antoine Godichon-Baggioni, Cathy Maugis-Rabusseau, and Andrea Rau

Abstract Although there is no shortage of clustering algorithms proposed in the literature, the question of the most relevant strategy for clustering compositional data (i.e., data made up of profiles, whose rows belong to the simplex) remains largely unexplored in cases where the observed value of an observation is equal or close to zero for one or more samples. This work is motivated by the analysis of two sets of compositional data, both focused on the categorization of profiles but arising from considerably different applications: (1) identifying groups of co-expressed genes from high-throughput RNA sequencing data, in which a given gene may be completely silent in one or more experimental conditions; and (2) finding patterns in the usage of stations over the course of one week in the Velib' bicycle sharing system in Paris, France. For both of these applications, we focus on the use of appropriately chosen data transformations, including the Centered Log Ratio and a novel extension we propose called the Log Centered Log Ratio, in conjunction with the K-means algorithm. We use a nonasymptotic penalized criterion, whose penalty is calibrated with the slope heuristics, to select the number of clusters present in the data. Finally, we illustrate the performance of this clustering strategy, which is implemented in the Bioconductor package coseq, on both the gene expression and bicycle sharing system data.

Keywords Clustering; compositional data; data transformations; K-means

Antoine Godichon-Baggioni
Sorbonne Université, France, e-mail: antoine.godichon_baggioni@upmc.fr

Cathy Maugis Rabusseau
INSA Toulouse, France, e-mail: maugis@insa-toulouse.fr

Andrea Rau
INRA Jouy en Josas, France, e-mail: andrea.rau@inra.fr

Entrepreneurial regimes classification: a symbolic polygonal clustering approach

Andrej Srakar, and Marilena Vecco

Abstract "Entrepreneurial regimes" is a topic, receiving quite a lot of research attention recently. As stated by some authors, for a more complete understanding of entrepreneurship contribution to economic and societal developments, it is important to recognize the contextually embedded quality of entrepreneurial actions, behaviors and frameworks in national, regional, and city-level contexts. 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, the entrepreneurial regimes is analyzed by applying a novel polygonal symbolic data cluster analysis approach and using Amadeus data for the period 2006-2015 covering 28 EU countries. Considering the diversity of data structures in Symbolic Data Analysis (SDA), interval-valued data is the most popular and many methods have been developed for this type of variable. This approach requires assuming equidistribution hypothesis. We use a novel polygonal cluster analysis approach to address this limitation which assumes the data are uniformly distributed in a polygon. The principal advantages of using polygonal data are: to store more information, to significantly reduce large data sets preserving the classical variability through polygon radius, independent of the number of objects into class, and to open new possibilities in symbolic data analysis.

Keywords symbolic data; polygonal cluster analysis; entrepreneurial regimes;

Andrej Srakar
Institute for Economic Research and University of Ljubljana, Slovenia, e-mail: andrej.srakar@ier.si

Marilena Vecco
CEREN, EA 7477, Burgundy School of Business, Université Bourgogne Franche-Comté, France, e-mail: marilena.vecco@bsb-education.com

Distances and discriminant analysis for microbial communities' composition to classify inflammatory bowel diseases

Glòria Mateu-Figueras, Pepus Daunis-i-Estadella, Mireia López-Siles, and Josep Antoni Martín-Fernández

Abstract The microbial community inhabiting the human intestine plays a fundamental role for health. A rising number of studies have reported that patients suffering intestinal disorders feature an altered gut microbiota, and this has been a topic systematically studied in inflammatory bowel diseases.

Using classical statistical analysis, it had been showed that Crohn's disease (CD) patients have an altered microbiota, which differs from that found in patients with ulcerative colitis (UC) and as well as of that in healthy controls. Although the location of the disease is a parameter relevant for clinical practice, few studies have explored changes in the microbial community among the subtypes of inflammatory bowel disease (IBD) that affect the colon (E1, E2, E3 and C-CD and IC-CD). Using a real data set, in this work we analyze the composition of the microbiota associated to the colonic mucosa of IBD patients. The aim is to identify bacterial markers that allow to discriminate diseases using distance discriminant analysis.

Distance and dissimilarities measures like Bray-Curtis or UniFrac are commonly used to compute the differences between microbial communities for non-parametric manova, dimensionality reduction or clustering methods. Recently the compositional nature of the microbiome data has been discussed recommending the methodology based on logratios. This opens the door to use the Aitchison distance as an adequate distance. The second aim of this work is to compare the performance of these measures for distance discriminant analysis using the previous real data set and other simulated data.

Keywords Gut microbiota; Aitchison distance; Bray-Curtis distance; UniFrac distance

Glòria Mateu-Figueras
University of Girona, Spain, e-mail: gloria.mateu@udg.edu

Pepus Daunis-i-Estadella
University of Girona, Spain, e-mail: pepus@imae.udg.edu

Mireia López-Siles
University of Girona, Spain, email: mireia.lopezs@udg.edu

Josep Antoni Martín-Fernández
University of Girona, Spain, e-mail: josepantoni.martin@udg.edu

Symbolic data analysis of gender-age-cause-specific mortality in European countries

Filipe Afonso, Aleša Lotrič Dolinar, Simona Korenjak-Černe, and Edwin Diday

Abstract Positioning of a country due to its mortality is important in order to implement proper health policies. Mortality data of European countries can be presented as aggregated data of gender-age-cause specific number of deaths. Analysis of such data requires the use of appropriate methods. Symbolic data analysis (SDA) offers tools for analysis of more complex data and can therefore also be used for aggregated data. We used methods of symbolic data analysis implemented in the SYR software to study the gender-age-cause specific mortality symbolic data of 28 European countries from the year 2015. The method shows clearly the behaviour contrast between the European countries.

The SYR software is used for the two-stage-automatic data processing of symbolic analysis: the first step involves fusion and reduction of the data into classes described by symbolic data. The tool is able to merge and aggregate heterogeneous data from multiple databases into a single symbolic data table. With this step, we obtained descriptions of our mortality data, i.e., histogram-valued representations of mortality levels and bar-charts of relative structure of mortality over cause of death. The second step is then the analysis of the resulting symbolic data by specific advanced statistical methods and machine learning techniques adapted for symbolic data presentations (e.g., dissimilarities, clustering, decision trees, factorial analysis, following paths based on metabins etc.). Our study showed that mortality level and cause are both discriminating (of the countries and of the clusters) for the different gender-age classes. Some countries have a discordant behavior. The program can automatically identify the most discriminating factors among clusters of countries. Our study also showed that groups of countries with similar behavior are related to the geographical position of countries.

Keywords symbolic data analysis; SYR program; mortality; causes of death

Filipe Afonso
Symbad – Le Symbolic Data Lab, Roissy CDG, France, e-mail: filipe.afonso@symbolicdata.com

Aleša Lotrič Dolinar
University of Ljubljana, School of Economics and Business, Slovenia, e-mail: alesa.lotric.dolinar@ef.uni-lj.si

Simona Korenjak-Černe
University of Ljubljana, School of Economics and Business, Slovenia, e-mail: simona.cerne@ef.uni-lj.si

Edwin Diday
CEREMADE, University Paris-Dauphine, France, e-mail: diday8@gmail.com

Clustering multivariate count data using a family of mixtures of multivariate Poisson log-normal distributions

Sanjeena Dang

Abstract Multivariate count data are commonly encountered through high-throughput sequencing technologies in bioinformatics. Although the Poisson distribution seems a natural fit to these count data, its multivariate extension is computationally expensive. Hence, independence between genes is assumed in most cases and this fails to take into account the correlation between genes. Recently, mixtures of multivariate Poisson lognormal (MPLN) models have been used to analyze these multivariate count measurements efficiently. In the MPLN model, the counts, conditional on the latent variable, are modeled using a Poisson distribution and the latent variable comes from a multivariate Gaussian distribution. Due to this hierarchical structure, the MPLN model can account for over-dispersion as opposed to the traditional Poisson distribution and allows for correlation between the variables. Here, a parsimonious family of mixtures of Poisson log-normal distributions is proposed by decomposing the covariance matrix and imposing constraints on these decompositions.

Keywords model-based clustering; multivariate count data

Sanjeena Dang
Binghamton University, USA, e-mail: sdang@binghamton.edu

Growth mixture modeling with measurement selection

Abby Flynt, and Nema Dean

Abstract Growth mixture models are an important tool for detecting group structure in repeated measures data. Unlike traditional clustering methods, they explicitly model the repeated measurements on observations, and the statistical framework they are based on allows for model selection methods to be used to select the number of clusters. However, the basic growth mixture model makes the assumption that all of the measurements in the data have grouping information that separate the clusters. In other clustering contexts, it has been shown that including non-clustering variables in clustering procedures can lead to poor estimation of the group structure both in terms of the number of clusters and cluster membership/parameters. In this talk, we will present an extension of the growth mixture model that allows for incorporation of stepwise variable selection based on the work done by Maugis, Celeux, and Martin-Magniette (2009) and Raftery and Dean (2006). Results presented on a simulation study suggest that the method performs well in correctly selecting the clustering variables and improves on recovery of the cluster structure compared with the basic growth mixture model. We also present an application of the model to a clinical study dataset and conclude with a discussion and suggestions for directions of future work in this area.

Keywords cluster analysis; growth mixture model; repeated measurements; longitudinal data; measurement selection

Abby Flynt
Bucknell University, USA email: abby.flynt@bucknell.edu

Nema Dean
University of Glasgow, Scotland, e-mail: Nema.Dean@glasgow.ac.uk

On the use of multiple scaled distributions for outlier detection and model-based learning

Brian Franczak, Antonio Punzo, and Cristina Tortora

Abstract Classification can be lucidly defined as the process of assigning group labels to sets of observations. When a finite mixture model is used for classification in either an unsupervised, semi-supervised, or supervised setting, one can refer to this process as model-based learning. In this talk, we will present a paradigm for parameterizing contamination and skewness within variants of the mixtures of shifted asymmetric Laplace (SAL) distributions. These models will be able to provide both group labels for like observations and detect whether an observation is an outlying point, unifying the fields of model-based learning and outlier detection. Of particular interest are the multiple scaled variants of the mixtures of SAL distributions which allow for directional contamination and skewness, resulting in contours that do not have the traditional elliptical shapes. Explicit details regarding the development of the proposed models will be provided and an expectation-maximization based parameter estimation scheme will be outlined. The classification performance of these models will be demonstrated using simulated and real data sets.

Keywords finite mixture models, shifted asymmetric Laplace, skewness

Brian Franczak
MacEwan University, Canada E-mail: franczakb@macewan.ca

Antonio Punzo
University of Catania, Italy E-mail: antonio.punzo@unict.it

Cristina Tortora
San Jose State University, CA, USA, e-mail: cristina.tortora@sjsu.edu

Skewed distributions or transformations? Accounting for skewness in cluster analysis

Michael P.B. Gallaugher, Paul D. McNicholas, Volodymyr Melnykov, and Xuwen Zhu

Abstract Due to its mathematical tractability, the Gaussian mixture model holds a special place in the literature. However, in a clustering scenario, using a Gaussian mixture model when skewness or outliers are present can be problematic. As a result, in recent years, many different methods have been proposed to account for skewed clusters. The two most prevalent methods in the literature are modelling skewness directly by using various skewed distributions for modelling the components and performing clustering alongside a suitable transformation. Although both these methods have been studied extensively in the literature and compared for select datasets in terms of relative performance, no extensive study has been performed to motivate in which situation to use one method over another. Moreover, in the case of using skewed distributions, which distribution to choose is also an ongoing question, as a more complex model may not always be the solution. Using many different real datasets, and looking at their underlying properties, such as measures of overlap between clusters, skewness, and kurtosis, we aim to provide more insight as to when one method – i.e., transformation or a skewed distribution – might be preferable to another. Simulated data and a large number of multivariate datasets will be considered.

Keywords clustering; mixture models; skewness; transformations

Michael P.B. Gallaugher
McMaster University, Canada, e-mail: gallaump@mcmaster.ca

Paul D. McNicholas
McMaster University, Canada, e-mail: paul@math.mcmaster.ca

Volodymyr Melnykov
University of Alabama, United States, e-mail: vmelnykov@cba.ua.edu

Xuwen Zhu
University of Louisville, United States, e-mail: xuwen.zhu@louisville.edu

Clustering of variables using CDPCA

Adelaide Freitas

Abstract Clustering and Disjoint Principal Component Analysis (CDPCA) is a constrained principal component analysis for multivariate numerical data aimed to identifying clusters of objects and, simultaneously, describing the multivariate data in terms of sparse and disjoint components. The set of variables that define each sparse and disjoint CDPCA component can be used to determine a cluster of variables. An alternating least-squares (ALS) algorithm was suggested to implement the CDPCA. Since this algorithm is iterative and requires an initialization step, it is crucial to evaluate whether (final) CDPCA components, and consequently, their correspondent clusterings of variables obtained from different applications of CDPCA on the same data set are similar. Considering several data sets (either the number of individual is lower or greater than the number of variables), we carried out an experimental comparative study in order to assess the performance of CDPCA based on ALS for providing variables clustering. The author was supported by Fundação para a Ciência e a Tecnologia (FCT), within project UID/MAT/04106/2019 (CIDMA- University of Aveiro).

Keywords clustering; PCA; sparse component

Adelaide Freitas
University of Aveiro, Portugal, e-mail: adelaide@ua.pt

A study of the variable outlyingness ranking that is obtained using different loading similarity coefficients

Sopiko Gvaladze, Kim De Roover, Francis Tuerlinckx, and Eva Ceulemans

Abstract In many research areas, studies result in multi-block data, in which the same variables are measured for different sets of observations (i.e., the data blocks are linked through the variable mode). Given such data, dimension reduction (DR) methods like principal component analysis (PCA) or factor analysis (FA) are highly popular, because they reduce the potentially large set of variables to a few constructs. DR methods yield score matrices that position the observations on the constructs, and loading matrices, linking the variables to the constructs. Very often researchers simply assume that the loadings are the same across the blocks. To check whether this assumption actually holds, multiple authors have proposed to compute some kind of similarity index (Tucker's congruence, Rv-coefficient, adjusted RV-coefficient, angle between subspaces, Root Mean Square Difference) between the loadings of the different blocks and investigate whether the similarity level is high enough. However, what to do when it is not? In many cases, this means that most of the variables behave comparably across the blocks, while some of them act differently. Therefore, one way to attain a more satisfactory similarity level is to remove the variables that affect the similarity level negatively. To do so, we propose to rank the variables according to their *outlyingness*. In the above, we already identified two possible choices – DR technique and similarity measure - that might potentially influence the correctness of the outlyingness ranking. We present the results of an extensive simulation study in order to investigate their effect in the two block case, revealing that using PCA and Tucker's congruence yields the best outlyingness rankings. We also illustrate the ideas by re-analyzing empirical data on sensory perceptions of different bread samples.

Keywords multivariate data; dimension reduction; principal component analysis; factor analysis; loading similarity; similarity coefficients; Tucker's congruence; Rv; angle between subspaces; outlyingness ranking

Sopiko Gvaladze
KU Leuven, Belgium, e-mail: sofia.gvaladze@kuleuven.be

Kim De Roover
Tilburg University, Netherlands, e-mail: K.DeRoover@tilburguniversity.edu

Francis Tuerlinckx
KU Leuven, Belgium, e-mail: francis.tuerlinckx@kuleuven.be

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@kuleuven.be

Some properties of coherent clusters of rank data

Vartan Choulakian

Abstract Let n voters rank order d items, and n>>d. Our aim is to find mixture components of rank data by taxicab correspondence analysis (TCA). The building block of a mixture component will be a coherent cluster. Essentially, we observe the following: On the first taxicab principal axis, the rank data is clustered into a finite number of clusters; a cluster can be of two kinds, coherent or incoherent. This talk will present some interesting mathematical properties of the coherent clusters, which will be applied on the SUSHI data set of size n=5000 and d=10. In particular, for interpretability we show that: first, a coherent cluster can be summarized-visualized by its contingency table of first-order marginals; second, Borda count rule can be used to provide a consensus ordering of the items representing the first taxicab principal dimension; third, crossing index measures the intermingling of scores of the voters between the optimal binary partition of the items and is independent of the sample size of the coherent cluster.

Keywords Borda count rule; contingency table of first-order marginals; crossing index

Vartan Choulakian
Université de Moncton, Canada, e-mail: vartan.choulakian@umoncton.ca

C443: A methodology to see a forest for the trees

Iven Van Mechelen, and Aniek Sies

Abstract Often tree-based accounts of statistical learning problems yield multiple decision trees which together constitute a forest. Reasons for this include examining tree instability, improving prediction accuracy, accounting for missingness in the data, and taking into account multiple outcome variables. A key disadvantage of forests, unlike individual decision trees, is their lack of transparency. Hence, an obvious challenge is whether it is possible to recover some of the insightfulness of individual trees from a forest. In this paper, we will briefly outline a conceptual framework and methodology to do so by reducing forests into one or a small number of summary trees, which may be used to gain insight into the central tendency as well as the heterogeneity of the forest. This is done by clustering the trees in the forest based on similarities between them. By means of a simulated data set we will demonstrate how and why different similarity types in the proposed methodology may lead to markedly different conclusions. We will finally illustrate the methodology with a study of tree instability by means of an empirical data set on the prediction of cocaine use on the basis of personality characteristics.

Keywords classification trees; statistical learning; bagging; ensemble methods; clustering

Iven Van Mechelen
University of Leuven, Belgium, email: iven.vanmechelen@kuleuven.be

Aniek Sies
University of Leuven, Belgium, email: aniek.sies@kuleuven.be

Assessing how feature selection and hyper-parameters influence optimal trees ensemble and random projection

Nosheen Faiz, Metodi Metodiev, Naz Gul, Andrew Harrison, Zardad Khan, and Berthold Lausen

Abstract This paper investigates the effect of feature selection on two proposed methods; Optimal Trees Ensemble and Random Projection in high dimensional settings. To this end, LASSO has been considered for selecting the most important features based on training data for dimension reduction. Additionally, the influence of various hyper-parameters regulating the two methods has also been assessed. Analysis on several benchmark datasets is given to illustrate the phenomena. The results reveal that feature selection improves the predictive performance of the Random Projection methods in addition to reducing the computational burden. The analyses also lead to practical directions of how to use these methods for the classification of big data.

Keywords optimal trees ensemble; random projection; high dimensional classification; feature selection

Nosheen Faiz
University of Essex, UK, e-mail: ne16298@essex.ac.uk

Metodi Metodiev
University of Essex, UK, e-mail: mmetod@essex.ac.uk

Naz Gul
Abdul Wali Khan University, Pakistan, email: sadeedalikhan@gmail.com

Andrew Harrison
University of Essex, UK, e-mail: harry@essex.ac.uk

Zardad Khan
Abdul Wali Khan University, Pakistan, email: zkhan@essex.ac.uk

Berthold Lausen
University of Essex, UK, e-mail: blausen@essex.ac.uk

Residual diagnostics for model-based trees for ordinal responses

Rosaria Simone, Carmela Cappelli, and Francesca Di Iorio

Abstract Among the consolidated methods to grow trees for ordinal responses, a prominent role is played by the model-based setting. This approach assumes a given parametric model (as the ordinal logit model) for each of the tree nodes. An alternative method is offered by CUBREMOT, a class of model-based trees for preference and evaluation data that is grounded on the specification of CUB models. In this setting, the application of residual diagnostics for ordinal data could enhance the understanding of the derived classification, its goodness, the best splitting criterion (if more are available), or help in tuning the tree depth for the post-pruning phase, for instance. Thus, also at a graphical level, residual analysis can support scholars in the choice of the best tree for given data. Summarizing, the proposal resorts to uniformity tests for the residuals built via a jittering approach to perform model selection for ordinal data. This issue will be investigated to assess the quality of model-based regression trees for rating responses. The trailhead of our analysis is a sub-sample of data taken from the 5th European Working Condition Survey carried out by Eurofound in 2010. We consider responses for Italy to the question` Do you experience stress in your work?'. The flexibility of CUBREMOT in shaping rating data at different subsetting levels, also in presence of some structural inflated categories, is enhanced by the residuals' diagnostic checks. Also for other case studies, results indicates that the combination of residual analysis with model-based trees can pave the way to broad applications and research in model selection and classification issues.

Keywords residuals; ordinal rating data; model-based trees

Rosaria Simone
Università degli Studi di Napoli Federico II, Italy, e-mail: rosaria.simone@unina.it

Carmela Cappelli
Università degli Studi di Napoli Federico II, Italy, e-mail: carcappe@unina.it

Francesca Di Iorio
Università degli Studi di Napoli Federico II, Italy, e-mail: fdiiorio@unina.it

Measuring and testing mutual dependence for functional data

Tomasz Górecki, Mirosław Krzyśko, and Waldemar Wołyński

Abstract We propose new measures of mutual dependence for multivariate functional data. Each measure is zero if and only if the vectors of functional features are mutually independent. The proposed measures base on the functional rV coefficient and distance correlation coefficient. The first one is appropriate for linear mutual dependence and the second one for non-linear mutual dependence between the vectors of functional features. Based on the proposed coefficients we can test mutual dependence. The implementation of corresponding tests is demonstrated by both simulation results and real data examples.

Keywords functional data; mutual dependence

Tomasz Górecki
Adam Mickiewicz University, Poland, e-mail: tomasz.gorecki@amu.edu.pl

Mirosław Krzyśko
Adam Mickiewicz University, Poland, e-mail: mkrzysko@amu.edu.pl

Waldemar Wołyński
Adam Mickiewicz University, Poland, e-mail: wolynski@amu.edu.pl

A co-clustering method for multivariate functional curves

Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Chèze, and Pauline Martin

Abstract The exponential growth of smart devices in all aspect of everyday life, leads to the collection of high frequency data for a same individual. Those data can be seen as functional data: a quantitative entity evolving along time for one individual. Connected devices also ease the collection of several variables simultaneously for the same individual, called multivariate functional data, which results in growing needs of methods to summarize and understand them. The example which has motivated this work is the monitoring of household electric consumption simultaneously with indoor and outdoor temperatures. In order to summarize such data, a clustering will produce homogeneous groups of individuals. Nevertheless, the interpretation of these clusters is difficult when the period of observation is long. Consequently, we propose to also summarize the temporal information by clustering days of observations which are similar. Resulting method is a co-clustering algorithm for a data matrix of multivariate functional data, which produces clusters of rows (individuals) as well as clusters of columns (days of observations). The crossing of rows and columns clusters leads to blocks of homogeneous multivariate functional observations.

The proposed approach relies on a functional latent block model, which assume for each block a probabilistic distribution for the scores of the multivariate curves resulting from a multivariate functional principal component analysis. Model inference relies on a SEM-Gibbs algorithm which alternates a SE-step where row and column partitions are simulated according to a Gibbs algorithm and a M-step where model parameters are updated thanks to the previous simulated partitions. To end, the best number of row and column clusters is selected thanks to the ICL criterion. The efficiency of the proposed algorithm will be illustrated on simulated data, then a practical example of smart houses will be analyzed.

Keywords co-clustering; multivariate functional data; smart houses

Amandine Schmutz
CWD-Vetlab, France, e-mail: aschmutz@lim-group.com

Julien Jacques
Université de Lyon, Lyon 2, France, e-mail: julien.jacques@univ-lyon2.fr

Charles Bouveyron
Université Côte d'Azur, France, e-mail: charles.bouveyron@unice.fr

Laurence Chèze
Université de Lyon, Lyon 1, France, e-mail: laurence.cheze@univ-lyon1.fr

Pauline Martin
Lim France, France, e-mail: pmartin@lim-group.com

One-way repeated measures ANOVA for functional data

Łukasz Smaga

Abstract In this paper, the one-way repeated measures analysis of variance for functional data is considered. For this problem, the pointwise test statistic is constructed by adapting the classical test statistic for the one-way repeated measures analysis of variance to functional data framework. The new test statistics are obtained by integrating and taking supremum of the pointwise test statistic. To approximate the null distributions of the test statistics and construct the testing procedures, different bootstrap and permutation methods are used. The performance of the new tests and their comparisons with the known testing procedures in terms of size control and power are established in simulation studies. These studies indicate that the new tests do not perform equally well and they are usually more powerful than the tests proposed in the literature. Illustrative real data example is also presented.

Keywords bootstrap; functional data; hypothesis testing; one-way ANOVA; permutation method; repeated measures

Łukasz Smaga
Adam Mickiewicz University, Poland, e-mail: ls@amu.edu.pl

Hidden Markov models for continuous multivariate data with missing responses

Fulvia Pennoni, Francesco Bartolucci, and Alessio Serafini

Abstract Hidden Markov models represent a popular tool for the analysis of longitudinal data, allowing the dynamic clustering of sample units on the basis of a set of repeated responses. In the literature on longitudinal data analysis, these models are typically used in the presence of multivariate categorical data, that is, when more categorical responses are observed at each time occasion. These formulations rely on the assumption of local independence, according to which the responses are conditionally independent given the latent states. Such assumption also simplifies the treatment of missing responses when the missing-at-random assumption is plausible. Here, we deal with the case of continuous multivariate responses in which, as in a Gaussian mixture models, it is natural to assume that the continuous responses for the same time occasion are correlated, according to a specific variance-covariance matrix, even conditionally on the latent states. Although maximum likelihood estimation of this model is straightforward in standard cases using the Expectation-Maximization algorithm, we focus on its estimation when: (*i*) suitable constraints on the variance-covariance matrix are assumed; (*ii*) there are missing responses. The constraints we refer to are commonly adopted in the literature of Gaussian finite mixture models. Regarding the assumptions on the generation of missing data we focus on the missing-at-random assumption and we also account for possible individual covariates that may directly affect the responses (in addition to the latent states). In particular, we propose an Expectation Maximization (EM) algorithm that provides exact maximum likelihood estimates and also computes standard errors for the parameter estimates. The proposed approach is illustrated by a simulation study, to evaluate the computational load, and through a real case analysis. We also show how the proposal may be useful in a context of time-series analysis with an application to financial data. An R implementation of the proposed algorithm is made available by the authors within the LMest package.

Keywords hierarchical clustering; expectation-maximization algorithm; forward-backward recursions; multivariate Gaussian distribution

Fulvia Pennoni
University of Milano-Bicocca, Italy, e-mail: fulvia.pennoni@unimib.it

Francesco Bartolucci
University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Alessio Serafini
University of Perugia, Italy, e-mail: alessio.serafini@unipg.it

Mixtures of cluster-weighted models with latent factor analyzer structure

Sanjeena Dang, Antonio Punzo, Salvatore Ingrassia, and Paul D. McNicholas

Abstract Cluster-weighted modeling (CWM) is a flexible statistical framework for modeling local relationships in heterogeneous populations using weighted combinations of local models. Cluster-weighted models are extended to include an underlying latent factor structure resulting in a family of Gaussian parsimonious cluster-weighted factor analyzers (CWFA) and a robust family of parsimonious cluster-weighted t-factor analyzers (CWtFA). In addition to the latent factor structure, CWtFA also contains the common factor analyzer structures. This provides even more parsimony, visualization options, and added flexibility when clustering high-dimensional data. The expectation-maximization framework along with Bayesian information criterion will be used for parameter estimation and model selection. The approach is illustrated on simulated data sets as well as a real data set.

Keywords model-based clustering; cluster-weighted models; high-dimensional data

Sanjeena Dang
Binghamton University, USA, e-mail: sdang@binghamton.edu

Antonio Punzo
University of Catania, Italy, e-mail: antonio.punzo@unict.it

Salvatore Ingrassia
University of Catania, Italy, e-mail: ingrassia@unict.it

Paul D. McNicholas
McMaster University, Canada, email: paul@math.mcmaster.ca

Specification of basis spacing for process convolution Gaussian process models

Herbert K. H. Lee, and Waley W. J. Liang

Abstract Gaussian process (GP) models have been widely used for statistical modeling of point-referenced data in many scientific applications, including regression, classification, and clustering problems. Standard specification of GP models is computationally inefficient for applications with a large sample size. One solution is to construct the GP by convolving a smoothing kernel with a discretized White noise process, which requires choosing the number of bases. The distance between adjacent bases plays a key role in model accuracy. In this paper, we perform a series of simulations to find a general rule for the basis spacing required for accurate representation of a discrete process convolution GP model. Under certain common conditions, we find that using a basis spacing of one-quarter the practical range of the process works well in practice.

Keywords Gaussian Processes; Process Convolutions; Spatial Modeling

Herbert K. H. Lee
University of California, Santa Cruz, e-mail: herbie@ucsc.edu

Waley W. J. Liang
University of California, Santa Cruz, e-mail: wliang@soe.ucsc.edu

Recursive partitioning of longitudinal and growth curve models

Marjolein Fokkema

Abstract Growth curve models are a popular tool to address to address questions of stability and change over time. Often, researchers may be interested in detecting subgroups which differ in initial levels, or in the direction or rate of growth over time. Several recursive partitioning methods allow for detecting subgroups in longitudinal data, like for example structural equation modeling trees (SEM trees), longRpart, RE-EM trees, linear model trees (LM trees) and linear mixed-effects model trees (LMM trees). All of these methods recursively partition the observations in a dataset into subgroups increasingly similar in terms of the response variable, while accounting for the dependence between observations over time through estimation of random effects. At the same time, the methods differ in terms of model specification and estimation procedures. This presentation provides and overview of the differences and similarities between the methods. Through a real data example on children's reading trajectories and a simulation study, the effects of different model specifications and estimation procedures will be evaluated. This will provide insights into how specific characteristics of the data problem can best be accounted for in recursive partitioning of longitudinal data and growth curve models.

Keywords recursive partitioning; growth curve models; longitudinal data

Marjolein Fokkema
Methodology and Statistics Unit Leiden University, The Netherlands, e-mail: m.fokkema@fsw.leidenuniv.nl

Bayesian regularization in probabilistic PCA with sparse weights matrix

Davide Vidotto

PCA allows detecting variables that correlate with each other within a specific latent variable representing the underlying dimension, also called component. The scores can be determined by using either the elements of the loadings matrix, or the elements of the weights matrix; the latter are useful to define each of the components.

Lately, penalized PCA methods relying on a lasso or elastic net approach have become prominent when analyzing high-dimensional data as these result in sparse solutions, this is solutions with many zero loadings or weights. As these kind of datasets tend to be composed of hundreds of thousands of variables, including many variables that are irrelevant for the latent construct considered, and are usually observed on a much smaller number of subjects, sparse approaches have important benefits: They may yield automatic selection of the relevant variables and are consistent in the high-dimensional setting. Furthermore, besides well-known methods such as matrix factorization and regression approaches, PCA can also be estimated via probabilistic modelling, by assuming a probability distribution for the observed data. This probability distribution determines the functional of the likelihood function. As a further step, placing a prior distribution upon the parameters of the likelihood leads to Bayesian PCA.

Bayesian PCA methods have shown promising results in the literature when sparsity is to be imposed on loading matrices. On the one hand, the addition of a prior distribution on the loadings allows achieving different degrees of regularization by means of hyperparameters manipulation. On the other hand, Bayesian analysis enjoys the unique feature that potential analyst's prior beliefs about the composition of the components can be automatically included into the model. While literature on Bayesian PCA has mainly focused on the loadings matrix, models that place priors on the components' weights have not been explored yet. With this talk, I am going to present a Bayesian PCA model where the focus is on the weights matrix. The model is estimated by means of a variational Bayes algorithm, which offers computational advantages w.r.t. Gibbs sampling. In order to achieve regularization under sparse conditions, a number of prior distributions for the variances of the components' weights is introduced. Moreover, approaches to variable selection are proposed that use posterior credibility intervals and stochastic variable selection (SVS) methods. Model's performance is investigated by means of a simulation study.

Keywords PCA; regularization; variational Bayes; dimension reduction

DavideVidotto
Tilburg University, The Netherlands, e-mail: d.vidotto@uvt.nl

Gaussian process panel modeling – statistical learning inspired analysis of longitudinal panel data

Julian Karch, Andreas Brandmaier, and Manuel Voelkle

Abstract To analyze longitudinal panel data, obtained from multiple individuals at multiple time points, a variety of psychometric modeling approaches, such as structural equation modeling, are used. These psychometric approaches are restricted to relatively simple models and rely on parametric statistical inference, which requires correct model specification. In contrast, statistical learning employs relatively complex models, and inferences do not require correctness of the model. To capitalize on these advantages of statistical learning, we extended the Bayesian nonparametric regression method Gaussian Process Regression for the analysis of longitudinal panel data. We termed this new approach Gaussian Process Panel Modeling (GPPM). GPPM provides enhanced flexibility both in terms of the models it can represent as well as the supported inference framework. In this talk, after introducing GPPM, we will focus on the utility of the hybrid models GPPM can represent. These hybrid models consist of a mix of parametric psychometric and nonparametric statistical learning models. When enough data is available, they profit from the flexible statistical learning component. Otherwise, they fall back to the restrictive psychometric component. As a result, they often outpredict both psychometric models and statistical learning models.

Keywords longitudinal data analysis; machine learning; statistical learning; Bayesian statistics; continuous-time modeling; prediction

Julian Karch
Leiden University, Netherlands, email: j.d.karch@fsw.leidenuniv.nl

Andreas Brandmaier
Max Planck Institute for Human Development, Germany, email: brandmaier@mpib-berlin.mpg.de

Manuel Voelkle
Humboldt University of Berlin, Germany, email: manuel.voelkle@hu-berlin.de

Finding the hidden link: sparse common component analysis

Katrijn Van Deun

Abstract Recent technological advances have made it possible to study human behavior by linking novel types of data to more traditional types of psychological data, for example linking psychological questionnaire data with genetic risk scores. Revealing the variables that are linked throughout these traditional and novel types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of depression. Little or no theory is available on the link between such traditional and novel types of data, the latter usually consisting of a huge number of variables. The challenge is to select - in an automated way - those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. To fill the methodological gap, we present here an extension of simultaneous component analysis. Constraints are introduced to impose block-structure and to force automated selection of the relevant variables. We will present an efficient procedure that is scalable to the setting of a very large number of variables. Using simulated data and an empirical example, we will showcase the benefits of the proposed method and compare with various competing methods, including sparse PCA and rotation techniques.

Keywords linked data analysis; simultaneous component analysis; regularization

Katrijn Van Deun
Tilburg University, Belgium, e-mail: k.vandeun@uvt.nl

Variants of three-way correspondence analysis: An R package

Rosaria Lombardo, Michel van de Velden, and Eric J. Beh

Abstract Four integrated R functions that perform variants of three-way correspondence analysis are examined for analysing bivariate and trivariate associations in three-way contingency tables with regard to nominal and ordered variables. The association in these contingency tables is modelled using correspondence analysis based on the generalised singular vectors of a three-mode component analysis (Tucker3) and on the polynomial components resulting from a trivariate moment decomposition. The package, called *CA3variants,* allows the user to choose from four variants of correspondence analysis that are applicable for the numerical and visual examination of association in three-way tables. These include the classical approach to three-way correspondence analysis, its non-symmetrical variant and the ordered symmetrical and non-symmetrical variants of three-way correspondence analysis. It also allows the analyst to consider the partitioning of the three-way indices on which the analysis are based upon. The package also provides tuning functions for selecting the model dimensionalities using the convex hull tool and/or bootstrap procedures.

Keywords three-way correspondence analysis; Pearson's statistic; Marcotorchino's index; bootstrap; convex hull

Rosaria Lombardo
University of Campania "Luigi Vanvitelli", Italy, e-mail: Rosaria.lombardo@unicampania.it

Michel van de Velden
University of Rotterdam, The Netherlands, e-mail: vandevelden@ese.eur.nl

Eric J. Beh
University of Newcastle, Newcastle, Australia, e-mail: eric.beh@newcastle.edu.au

Another view of Correspondence Analysis through Design and Projection matrices and General Linear Models

George Menexes, Angelos Markos, and Emmanouil D. Pratsinakis

Abstract In this study we present another view of Correspondence Analysis (CA) aiming at the exploration of the association between two categorical variables, especially in the case where one variable could be considered as dependent and the other as an independent one. The proposed methodological approach combines applications and attributes of Design and Projection "hat" matrices, properties of the Dual Scaling or the Homogeneity Analysis, the Principal of the Distributional Equivalence and aspects of the General Linear Models (especially MANOVA). Dual Scaling and Homogeneity Analysis are other names for Correspondence Analysis and constitute a class of methods aiming at the optimal quantification, under minimal restrictions, of the categories of the variables involved in the analysis. This quantification is optimal in the sense that some optimality criteria and mathematical-statistical properties are finally satisfied. Consequently, optimal scores could be derived for each sampling unit (subjects or objects) on each factorial axis resulted from CA. The design and the "hat" matrices are commonly used within the methodological frame of General Linear Models. According to the proposed method, initially we consider one of the two variables as the dependent and the other as the independent one. Next, the design matrix of the dependent variable is projected onto the (column) space of the independent variable. Following, we apply the Correspondence Analysis to appropriate matrices. The optimal quantified sampling units' scores could be analyzed further within the frame of MANOVA. A final "touch" of curve fitting exploration among the optimal values of the variables and some asymptotic results relative to the standard errors of the primary inertias (derived from Correspondence Analysis) are given. A basic result is that the proposed method could be applied to data resulted from planned experiments, where there is a clear distinction between dependent and independent variables.

Keywords dual scaling; homogeneity analysis; MANOVA

George Menexes
Aristotle University of Thessaloniki, Greece, e-mail: gmenexes@agro.auth.gr

Angelos Markos
Democritus University of Thrace, Greece, e-mail: amarkos@eled.duth.gr

Emmanouil D. Pratsinakis
Aristotle University of Thessaloniki, Greece, e-mail: pratsina@agro.auth.gr

Implicative and conjugative variables in the context of Correspondence Analysis

Odysseas Moschidis, and Angelos Markos

Abstract Correspondence Analysis (CA) is a method for multivariate data visualization designed to explore relationships among categorical (nominal) variables. In this paper, we extend its field of application to the case of implicative and conjugative variables, thanks to ingenious ways of recoding data to categorical scales. The joint analysis of simple, implicative and conjugative variables allows us to simplify and unveil latent structures and hidden relationships that lie in complex multivariate data sets.

Keywords correspondence analysis; complex data; interaction variables; data visualization

Odysseas Moschidis
University of Macedonia, Greece, e-mail: fmos@uom.gr

Angelos Markos
Democritus University of Thrace, Greece, e-mail: amarkos@eled.duth.gr

Combined use of Correspondence Analysis and Ordinary kriging to display "supplementary" values of quantitative variables onto the factorial planes

Thomas M. Koutsos and Georgios C. Menexes

Abstract In this work the combined use of Analyse Factorielle des Correspondances - AFC (or Correspondence Analysis) and the Ordinary Kriging method is proposed, as an effective way to display "supplementary" values of quantitative variables onto the factorial planes resulting from the application of AFC. The kriging is an effective spatial interpolation method that can be used for the estimation of a value at an un-sampled location-point by using a minimized estimation variance method, derived from a semi-variogram model, accounting for the spatial correlation of the neighbouring values. In the current study, a methodological scheme is proposed to display additional values of quantitative variables onto the factorial maps using hypothetical data from a 5×4 contingency table (sites×crops). Total amounts of fertilizers used were treated as additional information or "metadata". The proposed methodological scheme is aiming at a better interpretation of the graphical results of the AFC. Furthermore, this method can also be generalized in the case of Multiple Correspondence Analysis (Analyse des Correspondances Multiples).

Keywords supplementary points; spatial interpolation; factorial planes

Thomas M. Koutsos
Aristotle University of Thessaloniki, Greece, e-mail: tkoutsos@agro.auth.gr

Georgios C. Menexes
Aristotle University of Thessaloniki, Greece, e-mail: gmenexes@agro.auth.gr

Comparison of hierarchical clustering methods for binary data from SSR and ISSR molecular markers

Emmanouil D. Pratsinakis, Lefkothea Karapetsi, Symela Ntoanidou, Angelos Markos, Panagiotis Madesis, Ilias Eleftherohorinos, and George Menexes

Abstract Data from molecular markers, which are used to construct dendrograms based on genetic distances between different plant species, are encoded as binary data (0: absence of the band at the agarose gel, 1: presence of the band at the agarose gel). For the construction of the dendrograms, the most commonly used linkage method is the UPGMA (Unweighted Pair Group Method with Arithmetic mean) in combination with the squared Euclidean distance. It seems that in this scientific field, this is the 'golden standard' clustering method. In this study, a review is presented on the distances and the clustering methods used with binary data. Furthermore, an evaluation of the linkage methods (seven linkage methods) and the corresponding appropriate distances (27 distances) along with the combination of Benzécri's chi-squared distance with the Ward's linkage method, comparison of 189 clustering methods (except the 'golden standard') is attempted using data originating from molecular markers applied on various fruit trees species and *Sinapis arvensis* populations. Fruit trees cluster analysis was performed using SSR markers, while for *Sinapis arvensis* populations clustering were used ISSR markers. However, due to the nature of the SSR markers, it was observed that there were many "ties" in the distance matrix, that means too many samples had exactly the same proximities. The validation of the various cluster's solutions was tested using external criteria. The results showed that the 'golden standard' is not a 'panacea' for dendrogram construction, based on binary data derived from molecular markers. Thirty-seven other hierarchical clustering methods could be used for the construction of dendrograms from ISSR markers and eighty-seven other hierarchical clustering methods could be used for the construction of dendrograms from SSR markers.

Keywords dendrograms; proximities; linkage methods; Benzécri's chi-squared distance; correspondence analysis; categorical binary data; Sinapis arvensis; fruit trees

Emmanouil D. Pratsinakis
Aristotle University of Thessaloniki, Greece, e-mail: pratsina@agro.auth.gr

Lefkothea Karapetsi
CERTH, Centre for Research and Technology, Greece, e-mail: lkarapet@agro.auth.gr

Symela Ntoanidou
Aristotle University of Thessaloniki, Greece, e-mail: melina-nt@hotmail.com

Angelos Markos
Democritus University of Thrace, Greece, e-mail: amarkos@gmail.com

Panagiotis Madesis
Centre for Research and Technology, Greece, e-mail: pmadesis@certh.gr

Ilias Eleftherohorinos
Aristotle University of Thessaloniki, Greece, e-mail: eleftero@agro.auth.gr

George Menexes
Aristotle University of Thessaloniki, Greece, e-mail: gmenexes@agro.auth.gr

Inspecting smoking addiction of youth in Turkey through a latent class analysis

Ali Mertcan Köse, and Elif Çoker

Abstract Nowadays one of the most remarkable studies are about if smoking behaviour is a voluntary behaviour or is an addiction. The definition of addiction is given as *"The repeated involvement with a substance or activity, despite the substantial harm it now causes, because that involvement was (and may continue to be) pleasurable and/or valuable".* Although most of the smokers don't accept their addiction situation, in total there are 1.1 billion smokers in the world and around %80 of them live mostly in less economically developed countries. The average age of the smokers is 27.6 for Turkey. The increasing smoking behaviour led us to the purpose of the paper where the smoking behaviour is analyzed zor the students in Mimar Sinan Fine Arts University (MSGSU) in Istanbul, Turkey. The study consists of three different applications that are Confirmatory Factor Analysis (CFA), Latent Class Analysis (LCA) and ROC curve analysis conducted all together. In the application in order to measure the addiction of smoking, Cigarette Dependance Scale (CDS) -12 Likert scale is used. The CDS-12 scale questionnaire is applied to 651 students who are studying in MSGSU.

Primarily the reliability and validation analysis of CDS-12 scale is performed. Cronbach's Alpha value is found as 0.91 which can be interpreted as an excellent internal consistency. After checking the reliability and validity analysis, then CFA is applied. The CFA results suggested a very good fit (RMSEA=0.034, NFI=0.95, NNFI=0.97, CFI=0.98, GFI=0.99, AGFI=0.98). Next, LCA is used to determine the level of smoking addiction perception. The results of LCA suggested that smoking addiction of students can be categorized in three levels which can be named as low-level addicted (%23.5), middle-level addicted (49.7), high-level addicted (23.5). Moreover, the model has recommended significant results (AIC= 19860.94, BIC=20514.36, SSABIC=200050.808, LMR test ($p=0.0006<0.05$)). The entropy value is found as 0.904 which is a very satisfied result. Lastly, the low-level addicted and middle-level addicted classes are combined and ROC analysis is used to find the cut-off point for smoking addiction. By using the diagnose test of ROC curve, the cut-off point is found as 39. So, it means that if a student has a total score which is higher than 39, that student can be interpreted as a smoking addict. In conclusion, the 31% of the MSGSU students can be categorized as smoking addicts.

Keywords smoking behavior; cigarette addiction; confirmatory factor analysis; latent class analysis; roc curve

Ali Mertcan Köse
Mimar Sinan Fine Arts University, Turkey, e-mail: alimertcankose@gmail.com

Elif Çoker
Mimar Sinan Fine Arts University, Turkey, e-mail: elif.coker@msgsu.edu.tr

Data analysis on the annual use of the new deferasirox formulation in pediatric thalassemia patients

Symeon Symeonidis, Alkistis Adramerina, Aikaterini Teli, Nikoleta Printza, Antonios Papastergiopoulos, Labib Tarazi, Emmanouil Chatzipantelis, and Marina Economou

Abstract Thalassemia is the most common and severe chronic hemolytic anemia in Greece, the National Registry recently reporting more than 2000 transfusion-dependent patients. As a combined result of disease pathophysiology and transfusion treatment, thalassemic patients present with iron overload from an early age. Deferasirox (DFX), the newest of chelators used to excrete iron, was until recently administered in the form of dispersible tablets. However, the same active substance was licensed as film-coated tablets (FCT) to overcome issues related to drug intolerance. Because of the same active substance being used, no clinical trials in iron-overloaded patients were required from regulatory authorities. Aim of the present study was to evaluate the safety and efficacy of the new DFX formulation in pediatric patients with transfusion-dependent thalassemic syndromes, followed at a single pediatric center, were enrolled. Patients were prospectively evaluated as to hepatic function, glomerular and tubular renal function, and liver and cardiac iron overload over twelve months following initiation of DFX FCT. Statistical analysis was performed by SPSS 23 and MedCalc 14. Out of 19 patients enrolled 12 (63%) were males, and Mean patient age was 13.8 years (7-18 years). Before DFX FCT administration, 14 patients (73%) were on the previous DFX formulation (dispersible tablet, DT) and 5 patients on a different iron chelator. At the end of the study, mean ferritin values showed no statistically significant change. With regards to renal function, 18 patients (94.7%) presented with a reduction in glomerular filtration rate (GFR) at some time point during follow up. Transient increase of spot urinary protein: creatinine ratio was reported in 7 patients (36.8%), and of spot urinary calcium: creatinine ratio in 17 patients (89.5%) – in no case increase reaching statistical significance. Mild, transient elevation of alanine transaminase (ALT) and alkaline phosphatase was reported in 3 patients (15.7%) and of aspartate transaminase (AST) in 2 patients (10.5%). No statistically significant correlation was found between DFX FCT dose and laboratory parameters studied, i.e., ferritin, GFR, protein: creatinine and calcium: creatinine ratio, AST, ALT or ALP. All patients presented with a reduction in liver iron, which reached the limit of statistical significance, and a statistically significant reduction in cardiac iron. Reported drug-related adverse events were mild gastrointestinal symptoms in 3 patients (15.8%) and skin rash in 1 patient (5.2%). Study results demonstrated that the new DFX formulation is effective in reducing iron overload in thalassemic children while maintaining a satisfactory safety profile.

Keywords health data analysis; pediatric thalassemia; iron overload; deferasirox formulation

Symeon Symeonidis
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: ssymeoni@gmail.com

Alkistis Adramerina
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: alkistis_adrame@yahoo.com

Aikaterini Teli
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: katerina@med.auth.gr

Nikoleta Printza
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: nprintza@gmail.com

Antonios Papastergiopoulos
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: papastergiopoulos@med.auth.gr

Labib Tarazi
Tomografia AE, Medical Center, Thessaloniki, Greece, e-mail: tarazi@med.auth.gr

Emmanouil Chatzipantelis
2nd Pediatric Department, Aristotle University of Thessaloniki, University General Hospital of Thessaloniki AHEPA, Greece, e-mail: hatzip@auth.gr

Marina Economou
11st Pediatric Department, Aristotle University of Thessaloniki, Hippocrateion General Hospital of Thessaloniki, Greece, e-mail: marina@med.auth.gr

On missing label patterns in semi-supervised learning

Daniel Ahfock, and Geoffrey McLachlan

Abstract We investigate model based classification with partially labelled training data. The majority of theoretical work on semi-supervised learning makes the critical assumption that labels are missing uniformly at random. In many biostatistical applications, labels are manually assigned by experts, who may leave some observations unlabelled due to class uncertainty. Subjective labelling can lead to a non-uniform pattern of missing labels, and this has implications for likelihood-based inference. We analyse semi-supervised learning as a missing data problem and identify situations where the missing label pattern is non-ignorable for the purposes of maximum likelihood estimation. In particular, we find that a relationship between classification difficulty and the missing label pattern implies a non-ignorable missingness mechanism. We examine a number of real datasets and conclude the pattern of missing labels is related to the difficulty of classification. We propose a joint modelling strategy involving the observed data and the missing label mechanism to account for the systematic missing labels. Classification difficulty can be quantified using the Shannon entropy, and the subsequent influence on the labelling probability can be captured using a logistic selection model. Parameter estimation is feasible using profile likelihood methods. Full likelihood inference including the missing label mechanism can improve the efficiency of parameter estimation, and increase classification accuracy.

Keywords mixture models; missing data; semi-supervised learning

Daniel Ahfock
University of Queensland, Australia, e-mail: d.ahfock@uq.edu.au

Geoffrey McLachlan
University of Queensland, Australia, e-mail: g.mclachlan@uq.edu.au

Bayesian nonparametric mixture modeling for ordinal regression

Athanasios Kottas, and Maria DeYoreo

Abstract Univariate or multivariate ordinal responses are often assumed to arise from a latent continuous parametric distribution, with covariate effects which enter linearly. We will present Bayesian nonparametric methodology for univariate and multivariate ordinal regression, based on mixture modeling for the joint distribution of latent responses and covariates. The modeling framework enables highly flexible inference for ordinal regression relationships, avoiding assumptions of linearity or additivity in the covariate effects. In standard parametric ordinal regression models, computational challenges arise from identifiability constraints and estimation of parameters requiring nonstandard inferential techniques. A key feature of the methodology is that it achieves inferential flexibility, while avoiding these difficulties. In particular, the nonparametric mixture model has full support under fixed cut-off points that relate through discretization the latent continuous responses with the ordinal responses. The practical utility of the modeling approach will be illustrated through application to data sets from econometrics, an example involving regression relationships for ozone concentration, and a multirater agreement problem.

Keywords Dirichlet process mixtures; Markov chain Monte Carlo; Multivariate ordinal regression

Athanasios Kottas
University of California, Santa Cruz, CA, USA, e-mail: thanos@soe.ucsc.edu

Maria DeYoreo
RAND Corporation, Santa Monica, CA, USA, e-mail: mdeyoreo@rand.org

Assessment of recent social attitudes in Japan: a latent class item response theory model for web survey data

Miki Nakai, and Fulvia Pennoni

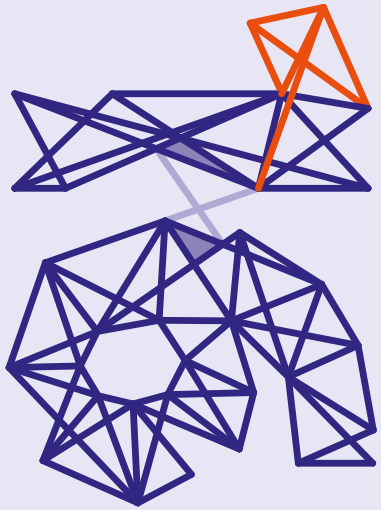
Abstract We illustrate a latent variable model included in the class of finite mixture models to analyze data from the Social Stratification and Social Psychology Project in Japan. The aim of the web-based survey carried out in December 2018 by an internet marketing research company is to identify and examine recent tendencies among the Japanese society by considering responses on social cognition and attitudes. It is important to understand the views that people hold about the social world and their evaluations of it. It is also important that how and whether diverse types of social attitudes differ by people's position in the society, such as social stratification position, generation, region, and gender. Survey requests were sent by email to target individuals chosen from a panel of more than 10 million members aged between 20 and 64 who have agreed to participate in online surveys and that are selected according to demographic quotas such as prefectural census population, age distribution, and gender. All participants received modest amount of voucher-based incentives for their time and effort.

We deal with a latent class-item response theory model for multivariate polytomous responses that is adapted to account for the web survey features. The model allows us to identify individual differences through the items of the questionnaire. Maximum likelihood estimation is performed through the Expectation-Maximization algorithm and the individuals are clustered by considering model selection principles such as BIC or the AIC. The model allows for predictions according with the maximum a-posteriori probability of the latent variable.

Keywords classification; Expectation Maximization algorithm; latent trait model; survey methodology

Miki Nakai
Ritsumeikan University, Japan, e-mail: mnakai@ss.ritsumei.ac.jp

Fulvia Pennoni
University of Milano-Bicocca, Italy, e-mail: fulvia.pennoni@unimib.it



POSTERS

The relationship of the apolipoprotein E genotype gene to the Alzheimer's disease: a meta-analysis

Sofia D. Anastasiadou

Abstract Meta-analysis aims to synthesize results of different studies with respect to the same subject and more specifically to the same research question. The data of the studies used are combined on the basis of a common metric such as the Odds Ratio (OR).

It has been confirmed by numerous studies that the Apolipoprotein E (APOE) gene is closely related to the occurrence of Alzheimer's Disease (AD) and is a stronger overall risk factor for the development of this disease.

This paper explores the effect of the Apolipoprotein E genotype (APOE) gene on Alzheimer's Disease (AD). The effect of the APOE gene was assessed with odds ratio (OR) at 95% confidence intervals (CIs).

Meta-analysis was performed to investigate the binding of the APOE gene to AD. The PubMed and Cochrane Library databases were used to find research on the relationship between APOE and Alzheimer Disease, which were included in Meta-analysis. The results showed the magnitude of this association.

Keywords meta-analysis; odds ratio; APOE; Alzheimer

Sofia D. Anastasiadou
University of Western Macedonia, Greece, e-mail: sofi.d.anastasiadou@gmail.com

Bayesian analysis for chromosomal interactions in hi-c data using hidden Markov random field model

Osuntoki G. Itunu, Andrew Harrison, Hongsheng Dai,Yanchun Bao, and Nicolae R. Zabet

Abstract There are few computational and statistical methods that have been developed over the years for data generated through the 3C-based methods, especially the Hi-C method. In our research, we develop a statistical methodology to explore the Hi-C data. We believe that the Hi-C data is well suited to be analysed using the finite mixture model, being a combination of background (noise) and signals. We make use of a hidden Markov model called Potts model.

The Potts model in our analysis is used to model the hidden (latent) components. Our hidden components are categorised into three; the Noise, the false signals and the True signals components. There are also some biases associated with Hi-C data that are incorporated as covariates into our model. Using the Metropolis-within-Gibbs approach to analysis our data, our method was able to detect the Topological Associated Domains (TADs) known to occur in Hi-C data. Our model was also able to detect long range interactions including false interactions.

Keywords Bayesian; hi-c; Metropolis-within-Gibbs; mixture model; Potts model

Osuntoki G. Itunu
University of Essex, United Kingdom, email: igosun@essex.ac.uk

Andrew Harrison
University of Essex, United Kingdom, email: harry@essex.ac.uk

Hongsheng Dai
University of Essex, United Kingdom, email: hdaia@essex.ac.uk

Yanchun Bao
University of Essex, United Kingdom, email: ybaoa@essex.ac.uk

Nicolae R. Zabet
University of Essex, United Kingdom, email: nzabet@essex.ac.uk

New financial instruments: Pollution emission rights and their trading on the stock exchange

Argiro Dimitoglou

Abstract In 2005, the Kyoto Protocol, which is an international agreement within the framework of the United Nations, is a key pillar for emissions control and the subsequent creation of a single European Emissions System. The Protocol sets emission rights, greenhouse gases, and CO2 emissions reduction.

The European Union (EU) and Greece, as its Member State, ratifying the Kyoto Protocol (Law 3017/2002), agreed to reduce anthropogenic greenhouse gas emissions in order to effectively protect the climate system. Recognizing climate change as a high priority, the EU has adopted Directive 2003/87 / EC establishing a well-functioning Community scheme for greenhouse gas emission allowance trading, aiming at a more effective fulfillment of its commitments by limiting as far as possible the negative effects on economic growth and employment. The Community Emissions Trading Scheme (ETS) is an operational Emission Trading System model which is constantly being upgraded and adopted by other countries.

At the initial stage of the EU-ETS, the allocation of allowances was made free of charge, and it is currently being auctioned on the primary and secondary markets with the introduction of rights on European stock exchanges. In Greece the auction of rights is made on the Athens Stock Exchange.

A typical example of integrated trading in the primary and secondary markets is the auctioning of pollution rights on the European Energy Exchange EEX, the most important stock exchange trading in Europe.

Keywords Kyoto Protocol; Pollutant Emissions Rights; European Emissions Scheme; Rights Exchange

Argiro Dimitoglou
University of Thessaly, Greece, e-mail: adimi@egnatia.gr

Econometric assessment of the relation between the situation of youth on the labour market and macroeconomic situation among the EU countries

Beata Bal-Domańska, and Elżbieta Sobczak

Abstract Many countries are experiencing serious problems connected with the entrance of young people into labour market. Young people, at the beginning of their professional career and often family life, face problems related to entering the labour market, obtaining "good" contracts, and adequate remuneration. The aim of the article is to assess the flexibility of the situation of young people in the labour markets depending on selected macroeconomic indicators and the characteristics of the labour market. The analysis will be carried out with the use of a workshop of econometric methods, including hierarchical clustering and regression models. Cluster analysis methods will identify groups of countries with both a similar macroeconomic situation and young people in the labour market. Econometric models will allow the description of the relationship between changes in the macroeconomic situation of a country (e.g. GDP, unemployment rate) and the employment of young people. The evaluation will be carried out among EU countries in the years 2003-2017.

Keywords hierarchical clustering; econometric modelling; youth on the labour market; EU member states

Beata Bal-Domańska
Wroclaw University of Economics, Poland, e-mail: beata.bal-domanska@ue.wroc.pl

Elżbieta Sobczak
Wroclaw University of Economics, Poland, e-mail : elzbieta.sobczak@ue.wroc.pl

Comparison of pattering methods: Clustering of variables, Implicative Statistical Analysis and Analysee Factorielle des Correspondances

Sofia D. Anastasiadou

Abstract The current paper intends to gain insight about the outcomes, common or different, advantages and contribution of three particular pattering methods named Clustering of variables, Implicative Statistical Analysis and Analysee Factorielle des Correspondances by putting side by side the results of their application in evaluating the understanding and learning of probabilities notions. These notions have different modes of representations, such as verbal, algebraic and graphical. Data were obtained from 238 students from pedagogical departments. It is of a major importance in statistics education the evaluation of students' abilities in transportations and conversions from a representation mode to another.

Results were shown to be comparable in relation to probabilities notions' understanding. In addition, they showed that the three methods operate complementary, each one accentuating a different dimension for the interpretation of data, the interpretation of which presents a different aspect in students' capability in learning probabilities notions.

Keywords clustering of variables; implicative statistical analysis; analysee factorielle des correspondances

Sofia D. Anastasiadou
University of Western Macedonia, Greece, e-mail: sofi.d.anastasiadou@gmail.com

Framing coworking spaces digital marketing strategy via social media analytics

Dimitrios Vagianos, and Nikos Koutsoupias

Abstract The emergence of Coworking Spaces as the workplace of the future has been inspirational for forming and analyzing datasets in this paper, as data have been collected utilizing Social Media Monitoring techniques related to digital marketing. Having focused on a case study of a coworking company, we use Mediatoolkit Social Media marketing tool to collect data derived from the activity of the company's Instagram and Twitter accounts were on a 24/7 basis from varying locations and in multiple languages in a fifteen days' time span. Indices related to sentiment, reach, influence, number of followers, retweets, likes, comments and view scores formed the datasets that are explored using both Multiple Factor Analysis and Hierarchical Clustering.

The analysis in this paper attempts to investigate the inherent properties of the multiple indices describing the general realm of Social Media Marketing tools and more specifically aspires to provide digital marketers with an alternative perspective of social media marketing strategies related to the Coworking Spaces.

Keywords multiple factor analysis; hierarchical clustering; social media analytics, coworking spaces

Nikos Koutsoupias
University of Macedonia, Greece, e-mail: nk@uom.gr

Dimitrios Vagianos
University of Macedonia, Greece, e-mail: vagianos@uom.gr

Sales performance measure: A systematic review and typology of research studies

Tor Korneliussen, Per Seljeseth, and Michael Greenacre

Abstract Valid and precise measures of sales performance outcomes are fundamental for knowledge building in sales research. However, the literature provides little consensus or guidelines on which measures are appropriate for assessing sales performance outcomes. This study contributes with a systematic review of the measures that researchers use for assessing sales performance outcomes in business-to-business (B2B) selling. The review of the 2001-2015 period identified 139 studies published in 17 journals, revealing that researchers have used a large variety of 151 measures. A cluster analysis provides a typology of seven groups of studies with homogeneous measures. The methodological challenge in this study is to deal with a large very sparse data table.

Keywords sales performance measures; hierarchical clustering; sparse tables

Tor Korneliussen
Nord University Business School, Norway, email: tor.a.korneliussen@nord.no

Per Seljeseth
Nord University Business School, Norway, email: per.i.seljeseth@nord.no

Michael Greenacre
Universitat Pompeu Fabra and Barcelona Graduate School of Economics, Spain, email: michael.greenacre@upf.edu

Document clustering via multiple correspondence, term and metadata analysis in R

Nikos Koutsoupias, and Kyriakos Mikelis

Abstract We introduce the combined use of multiple correspondence analysis, metadata and term frequencies for clustering articles of a scientific journal. A period of five years (2010-2014) is covered, with approximately 125 articles. Through specific R packages for multidimensional data analysis and text mining, the approach links quantitative analysis of discourse to clustering documents considering both metadata and frequent terms.

Keywords document clustering; hierarchical clustering; multiple correspondence analysis; document metadata; text mining

Nikos Koutsoupias
University of Macedonia, Greece, e-mail: nk@uom.gr

Kyriakos Mikelis
University of Macedonia, Greece, e-mail: mikelis@uom.gr

Comparison of multivariate methods in group/cluster identification: PCA vs discriminant analysis and K-Means clustering

Sofia D. Anastasiadou

Abstract Even though there is a substantial development and utilization of pattering methods in medicine, a direct comparison of multivariate methods in group/cluster identification for biomarkers has not been carried out. This paper analyses and compares three different statistical techniques: i.e the Principal Components Analysis (PCA), the Discriminant Analysis and the K-Means clustering with respect to biochemical measurements.

The study included 303 patients, 151 cases and 152 controls. The 151 patients (cases) were diagnosed as suffering from kidney disease. Concentrations of AST (SGOT), ALT (SGPT), Glucose Serum, Urea, Creatinine, Serum Uric Acid, Serum Calcium, Potassium Serum, Sodium Serum, Total Albumins (TP), Albumin, Alp, γ-GT, CRP, LDH and CPK were measured.

PCA's results showed the existence of 5 Components, amongst which the third is shown to be the Component for renal function. This Component comprises of variables: Urea, Creatinine and Serum Uric Acid, which are also the variables which are clinically measured to determine the existence or not of kidney disease. From the scatter plots for all combinations of Components, it was established that the Component for renal function was indeed the one with respect to which controls differentiated from cases. Discriminant Analysis was applied twice. It was initially applied on all 16 variables measuring the concentrations in the participants' biochemical analyses and showed that Urea is indeed the best predictor, followed by Creatinine and then Serum Uric Acid, all with respect to separating controls from cases. The accuracy of Predicted Group Membership was verified. Moreover, analysis exhibits high sensitivity and high specificity. It was then applied only for aforementioned three variables and showed that they are, indeed, the appropriate predictors for the separation of the two groups, controls from cases. More specifically, Creatinine was shown to be the best predictor, followed by Urea and Serum Uric Acid, with respect to the separation of controls from cases. Predicted Group Membership accuracy was verified in this analysis as well, as were the high sensitivity and high specificity of the data.

K-Means was applied only on these three variables and showed that Urea predictor, Creatinine and Serum Uric Acid predictors can satisfactorily separate controls from cases. Results were shown to be comparable in relation to plasma biomarkers and kidney disease.

Keywords PCA; Discriminant Analysis; K-Means; Clustering

Sofia D. Anastasiadou
University of Western Macedonia, Greece, e-mail: sofi.d.anastasiadou@gmail.com

Asymptotic cumulants of the minimum phi-divergence estimator for categorical data under possible model misspecification

Haruhiko Ogasawara

Abstract The asymptotic cumulants of the minimum phi-divergence estimators of the parameters in a model for categorical data are obtained up to the fourth order with the higher-order asymptotic variance under possible model misspecification. The corresponding asymptotic cumulants up to the third order for the studentized minimum phi-divergence estimator are also derived. These asymptotic cumulants, when a model is misspecified, depend on the form of the phi-divergence. Numerical illustrations with simulations are given for typical cases of the phi-divergence, where the maximum likelihood estimator does not necessarily give best results. Real data examples are shown using log-linear models for contingency tables.

Keywords asymptotic variance; bias; skewness; studentization; log-linear models

Haruhiko Ogasawara
Otaru University of Commerce, Japan, e-mail: emt-hogasa@emt.otaru-uc.ac.jp

Multidimensional data analysis in perception of European Union by different generations

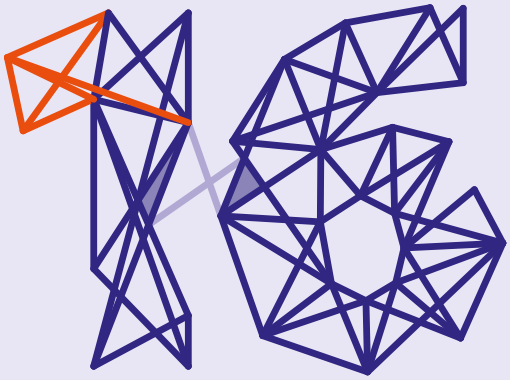
Agnieszka Stanimir

Abstract In the presentation the results of analysis carried out on the perception of socio-economic solutions proposed by the European Union for various generations will be presents. The aims of the analyses are changes in time of EU functioning perception. The studies were also focused on indication of adult generations of Europeans assessment of EU and how they profit the solutions proposed by the Union. The evaluations of the functioning of the EU and the programs proposed by the Union were presented in relation to the assessment of the quality of life, both in subjective terms and by comparing the values of objective indicators. The variables used in the study are measured on non-metric scales. Therefore, adequate, and various analytical methods were used to carry out the analyses, so as to describe the research problem in the largest, possible way.

Keywords generational differences; perception of EU; quality of life; multidimensional data analysis; non-metric data

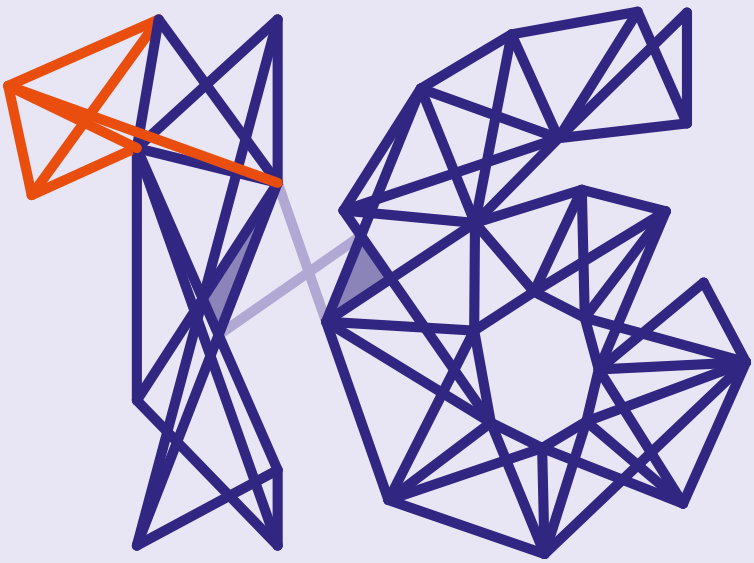
Agnieszka Stanimir
Wroclaw University of Economics, Poland, e-mail: agnieszka.stanimir@ue.wroc.pl

INDEX



INDEX

Z



WITH THE SUPPORT OF

HELLENIC STATISTICAL AUTHORITY

THANK YOU FOR YOUR SUPPORT!

WELCOME RECEPTION SPONSOR

ΕΚΔΟΣΕΙΣ
GUTENBERG

CONFERENCE BAG SPONSOR

Εκδόσεις
Αθανασίου Αλτιντζή

SPONSORS

Επιστημονικές Εκδόσεις
ΤΖΙΟΛΑ

ΕΠΙΣΤΗΜΟΝΙΚΕΣ ΕΚΔΟΣΕΙΣ
ΠΑΡΙΣΙΑΝΟΥ Α.Ε.
www.parisianou.gr • medbooks@parisianou.gr

Publications
PROPOBOS
www.propobos.gr

COMMUNICATION SPONSOR

DatAnalysis
STATISTICAL EXCELLENCE

16th Conference
of the International
Federation of
Classification Societies

ARTION
conferences & events
PROFESSIONAL CONGRESS ORGANISER
FOR IFCS-2019 CONFERENCE
E. ifcs@artion.com.gr
T. +30 2310 257803 (direct line), +30 2310 272275