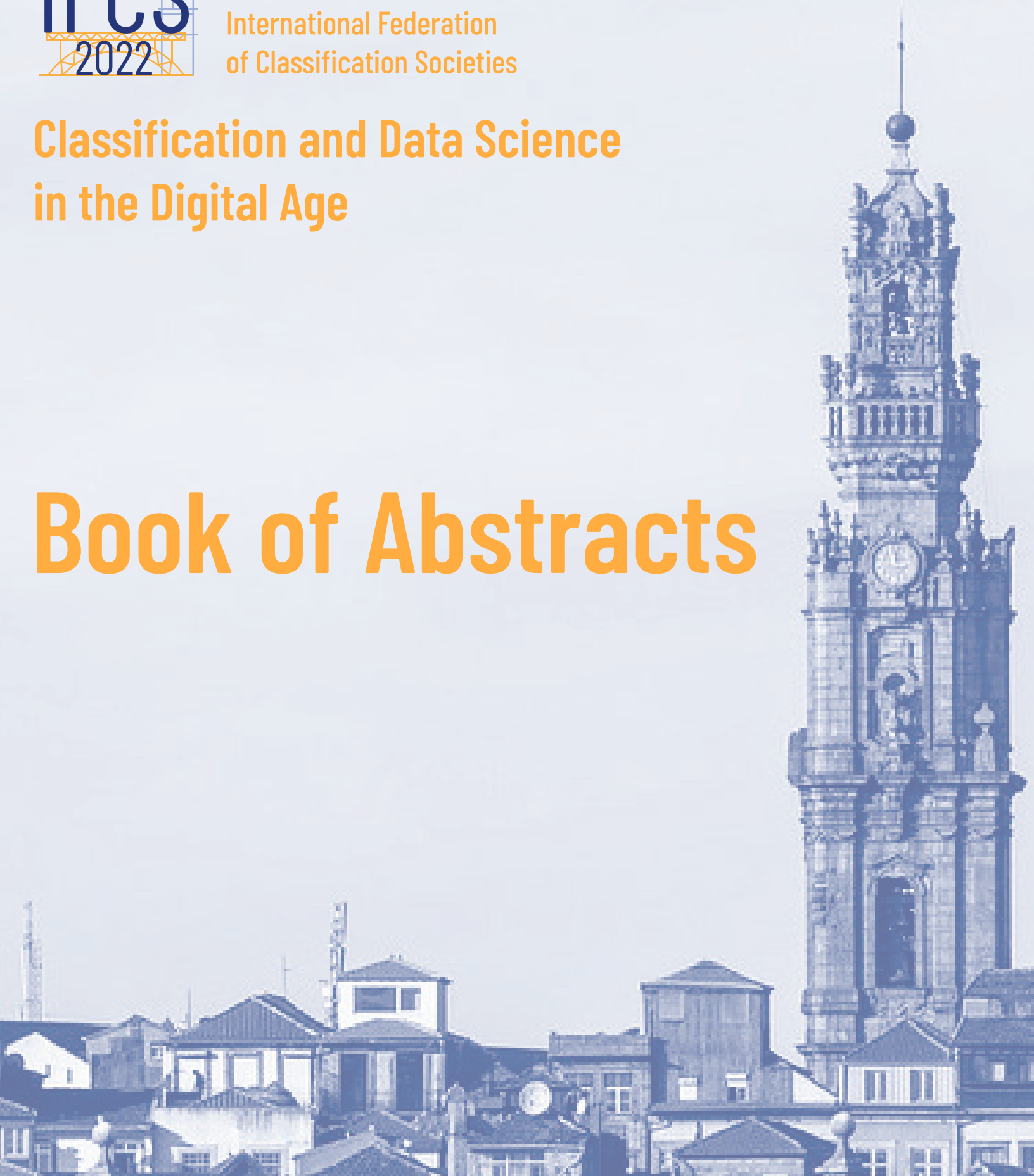
IFCS 2022

17th Conference of the International Federation of Classification Societies

Classification and Data Science in the Digital Age

Book of Abstracts



IFCS 2022

Book of Abstracts

17th Conference of the International Federation of
Classification Societies

Classification and Data Science in the Digital Age

PORTO, PORTUGAL

Title: Classification and Data Science in the Digital Age - Book of Abstracts IFCS 2022
Authors: CLAD - Associação Portuguesa de Classificação e Análise de Dados
Cover design: exclamação!
Printed in Portugal by Instituto Nacional de Estatística
ISBN: 978-989-98955-9-1
N. DL: 500335/22
Number of copies: 350

Preface

Welcome to the 17th conference of the IFCS, IFCS 2022, held in Porto, Portugal, from July 19 to July 23, 2022 and the first IFCS conference held in Portugal. It is a joint organisation of the Portuguese Association for Classification and Data Analysis, CLAD, and the Faculty of Economics of the University of Porto, FEP-UP.

IFCS 2022 is preceeded by two half-day tutorials, one on *Analysis of Data Streams* by João Gama, and another on *Categorical Data Analysis and Visualization* by Rosaria Lombardo and Eric J Beh, features four keynote speakers, five invited and seventy contributed sessions organised in specific topics. The Benchmarking Challenge, the Awards Session and the Presidential Address are also noteworthy. Overall, the call for papers attracted 280 submissions, representing 42 countries and 578 different authors. The authors come from five continents, being the largest representation from Europe (68%), followed by North America (12%). Additionally to the rich scientific program the LOC has organised a number of social appealling events that will be memorable.

The 17th conference of the IFCS would not have been possible without the support of many individuals and organisations. We owe special thanks to the authors of all the submitted papers, the members of the program committee, and the reviewers for their contributions to the success of the conference. Finally, we acknowledge the institutional and industrial sponsors that contributed to the organisation of the conference. In particular, we thank all those at FEP-UP who enthusiastically supported the conference from the very start, contributing to its success.

This book contains the abstracts corresponding to all the presentations at the conference. It is organised in seven parts, according to the type of session. Within each session the abstracts are ordered according to the programme. The book includes also an author index.

It has been a pleasure and an honor to organise and host IFCS 2022 in Porto. It is our wish that all participants enjoy the scientific program as well as the the social events and the city of Porto and Portugal.

July 2022 The Local Organising Committee

Organisation

Scientific Programme Committee

Ex-Officio

Paula Brito	Co-Chair	Universidade do Porto, Portugal
José G. Dias	Co-Chair	ISCTE- Instituto Universitário de Lisboa
Angela Montanari	IFCS President	Università di Bologna, Italy
Berthold Lausen	IFCS Past President	University of Essex, UK

Representatives of IFCS Member Societies

Theodore Chadjipadelis	GSDA
Brian Franczak	CSNA
Krzysztof Jajuga	SKAD
Hyunjoong Kim	KCS
Simona Korenjak-Černe	SSS
Koji Kurihara	JCS
Francesco Mola	CLADAG
Ahmed Moussa	MCSo
Fionn Murtagh	BCS
Mohamed Nadif	SFC
Mark de Rooij	VOC
Niël le Roux	MDAG
Eva Boj del Val	SEIO
Arthur White	IPRCS
Adalbert Wilhelm	GfKl
Javier Trejos Zelaya	SoCCCAD

Additional Members

Agustín Mayo-Íscar	Geoffrey McLachlan
André Carvalho	Glòria Mateu-Figueras
Anuska Ferligoj	Hans Kestler
Carlos Soares	José Antonio Vilar Fernández
Cinzia Viroli	Karell Bertet
Eyke Hüllermeier	Laura M. Sangalli
Francisco de Carvalho	Laura Palagi

Lazhar Labiod
Luis Ángel García Escudero
Panduranga Nagabhushan
Pedro Campos
Peter Filzmoser
Rosanna Verde
Rosaria Lombardo
Salvatore Ingrassia
Satish Singh
Sibel Kazak
Veronica Piccialli
Vladimir Batagelj

Reviewers

Adalbert Wilhelm
Agustín Mayo-Íscar
Alípio Jorge
André C. P. L. F. de Carvalho
Ann Maharaj
Anuška Ferligoj
Arthur White
Berthold Lausen
Brian Franczak
Carlos Soares
Christian Hennig
Conceição Amado
Eva Boj del Val
Francesco Mola
Francisco de Carvalho
Geoff McLAchlan
Gilbert Saporta
Glòria Mateu-Figueras
Hans Kestler
Hélder Oliveira
Hyunjoong Kim
Jaime Cardoso
Javier Trejos
Jean Diatta
José A. Lozano
José A. Vilar
José Matos
Koji Kurihara
Krzysztof Jajuga
Laura Palagi
Laura Sangalli
Lazhar Labiod
Luis Ángel García-Escudero
Luis Teixeira
M. Rosário Oliveira
Margarida G. M. S. Cardoso
Mark de Rooij
Michelangelo Ceci
Mohamed Nadif
Niël Le Roux
Paolo Mignone
Patrice Bertrand
Pedro Campos
Pedro Duarte Silva
Pedro Ribeiro
Peter Filzmoser
Rosanna Verde
Rosaria Lombardo
Salvatore Ingrassia
Satish Singh
Simona Korenjak-Černe
Theodore Chadjipadelis
Veronica Piccialli
Vladimir Batagelj

Local Organising Committee

Paula Brito (Chair)	Universidade do Porto
Adelaide Figueiredo	Universidade do Porto
Carlos Abreu Ferreira	Instituto Politécnico do Porto
Carlos Marcelo	INE
Conceição Rocha	INESC TEC
Fernanda Figueiredo	Universidade do Porto
Fernanda Sousa	Universidade do Porto
Jorge Pereira	Universidade do Porto
Maria Eduarda Silva	Universidade do Porto
Paulo Teles	Universidade do Porto
Pedro Campos	Universidade do Porto
Pedro Duarte Silva	Universidade Católica - Porto
Sónia Dias	Instituto Politécnico de Viana do Castelo

Partners& Sponsors

We are extremely grateful to the following institutions whose support contributes to the success of IFCS 2022:

Sponsors

Partners

Organisation

Contents

Part I Presidential Address

Perturb and Conquer: how Classification Can Benefit from Data Perturbation 2
Angela Montanari

Part II IFCS Research Medal Lecture

An Introduction to S-concordance and S-discordance 4
Edwin Diday

Part III Tutorials

Analysis of Data Streams 6
João Gama

Categorical Data Analysis and Visualization 7
Rosaria Lombardo and Eric J. Beh

Part IV Keynote Lectures

A Showcase of New Methods for High Dimensional Data Viewing with Linear Projections and Sections 10
Dianne Cook

Statistical Learning with Dynamic Interaction Data for Public Health . . . 11
Charles Bouveyron

Fast Minipatch Ensemble Strategies for Learning and Inference 12
Genevera I. Allen

Trends in Data Stream Mining 13
João Gama

Part V Invited Sessions

A Criterion for Selecting the Number of Time Series Clusters 16
Daniel Peña and Ruey S. Tsay

Clusters Based on Prediction Accuracy of Global Forecasting Models . . . 17
Pablo Montero-Manso, Ángel López-Oriona, and José A. Vilar

Clustering and Classifying Time Series in the Sktime Toolkit: a Practical Review of Latest Advances in the Field 18
Anthony Bagnall

Effect of Type 2 Diabetes and its Genetic Susceptibility on Severity and Mortality of COVID-19 in UK Biobank 19
Aeyeon Lee, Youngkwang Cho, Jun Li, Taesung Park, Wonil Chung, and Liming Liang

Prognosis of COVID-19 Patients by the Underlying Diseases and Drug Treatment in Korea 20
Ho Kim and Taerim Lee

Algorithms for Clustering COVID-19 Data: a Holistic Overview of Current Trends and New Visual Approaches 21
Eun-Kyung Lee

Embedded Word MCA Biplots for Sentiment Visualisation: Application to COVID-19 Related Tweets 22
Zoë-Mae Adams, Johané Nienkemper-Swanepoel, Niël le Roux, and Sugnet Lubbe

A General Framework for Implementing Distance Measures for Categorical Variables 23
Carlo Cavicchia, Michel van de Velden, Alfonso Iodice D'Enza, and Angelos Markos

Some Descriptive Statistics of Aggregated Symbolic Data 24
Junji Nakano, Nobuo Shimizu, and Yoshikazu Yamamoto

Variable Screening in High Dimensional Regression via Random Projection Ensembles 25
Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, and Angela Montanari

Model-based Clustering and Dimension Reduction for Multidimensional Social Networks 26
Michael Fop, Silvia D'Angelo, and Marco Alfò

Conditional Gaussian Mixture Modeling 27
Volodymyr Melnykov and Yang Wang

Classification Over Text, Relational Databases and Graphs - Software and Case Studies 28
Tomáš Kliegr

Towards Deep and Interpretable Rule Learning 29
Johannes Fürnkranz

Current Challenges in Interpretable Machine Learning and Partitioning Approaches 30
Bernd Bischl

Part VI Benchmarking Challenge

Vine Copula Mixture Models and Clustering for Non-Gaussian Data 32
Özge Sahin and Claudia Czado

Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering 33
Zdeněk Šulc and Hana Řezanková

Comparing Model Selection Techniques to Determine the Number of Overlapping Clusters for the Additive Profile Clustering Model 34
Tom F. Wilderjans, Julian Rossbroich, and Jeffrey Durieux

Pitfalls of Automatic Optimization Procedures and Benchmarking in Cluster Analysis 35
Quirin Stier and Michael C. Thrun

Part VII Contributed Abstracts

Biplots for Categorical and Ordinal Data Based on Logistic Responses . . 38
Jose Luis Vicente-Villardon

The Biplot Inner Product for Interpretation and Derivation of Eigenvector Methods 39
Cajo J. F. ter Braak

Fifty Years of Biplots: Some Remaining Enigmas and Challenges 40
Jan Graffelman

Outlier Detection for BIG Functional Data 41
Rosa E. Lillo, Oluwasegun T. Ojo, and Antonio Fernández-Anta

Outlier and Novelty Detection for Functional Data: a Semiparametric Bayesian Approach 42
Francesco Denti, Andrea Cappozzo, and Francesca Greselin

A Geometric Perspective on Functional Outlier Detection 43
Moritz Herrmann and Fabian Scheipl

A New Decomposition of Orthogonal Matrices with Application to Common Principal Components 44
Luca Bagnato and Antonio Punzo

An MML Embedded Approach for Estimating the Number of Clusters . . 45
Cláudia Silvestre, Margarida G. M. S. Cardoso, and Mário Figueiredo

Comparison of Segmentation Approaches for Partial Least Squares Path Modeling with Stability Assessment 46
Sophie Dominique, Mohamed Hanafi, Fabien Llobell, Jean-Marc Ferrandi, and Véronique Cariou

Robust Classification for Toroidal Data 47
Giovanni Saraceno, Luca Greco, and Claudio Agostinelli

Consistency of Trimmed Estimators of Scatter Under the t-distribution . . 48
Andrea Cerioli, Lucio Barabesi, Luis A. García-Escudero, and Agustín Mayo-Iscar

Robust Classification in High Dimensions Using Regularized Covariance Estimates 49
Valentin Todorov and Peter Filzmoser

Symbolic Concordance and Discordance Illustrated on Data from an International Teaching and Learning Survey 50
Simona Korenjak-Černe, Barbara Japelj Pavešić, and Edwin Diday

A Clusterwise Regression Method for Distributional Data 51
Rosanna Verde, Antonio Balzanella, and Antonio Irpino

The Use of Regression to Partition a Dataset of Interval Observations . . . 52
Lynne Billard and Fei Liu

Heterogeneous Random Forests 53
Ye-eun Kim and Hyunjoong Kim

Analysis of the Damage Rate Using Typhoon Information 54
Su Hoon Choi and Min Soo Kim

Resampling, Relabeling, and Raking for Extremely Imbalanced Classification 55
Hae-Hwan Lee, Seunghwan Park, and Jongho Im

Clustering Student Mobility Data in 3-way Networks 56
Vincenzo Giuseppe Genova, Giuseppe Giordano, Giancarlo Ragozini, and Maria Prosperina Vitale

Multi-perspective Risky User Classification in Social Networks 57
Antonio Pellicani, Gianvito Pio, and Michelangelo Ceci

Clustering and Blockmodeling Temporal Networks – Two Indirect Approaches 58
Vladimir Batagelj

Clustering Validation in Hierarchical Cluster Analysis: an Empirical Study 59
Osvaldo Silva, Áurea Sousa, and Helena Bacelar-Nicolau

Divide and Conquer: a Clustering Method for Hierarchical and Nested Data Structures 60
Andrej Svetlošák, Miguel de Carvalho, Gabriel Martos Venturini, and Raffaella Calabrese

Significance Mode Analysis (SigMA) for Hierarchical Structures 61
Sebastian Ratzenböck, Torsten Möller, Josefa E. Großschedl, João Alves, Immanuel M. Bomze, and Stefan Meingast

Kurtosis-based Projection Pursuit for Matrix-valued Data 62
Una Radojicic, Klaus Nordhausen, and Joni Virta

Comparison of Pixel Based Segmentation Methods in Papillary Thyroid US Images 63
Neslihan Gökmen İnan, İsmail Meşe, Düzgün Yıldırım, and Ozan Kocadağlı

Bootstrapping Binary GEV Regressions for Massive Unbalanced Datasets 64
Michele La Rocca, Marcella Niglio, and Marialuisa Restaino

Stochastic Collapsed Variational Inference for Structured Gaussian Process Regression Networks 65
Rui Meng, Herbert K. H. Lee, and Kristofer Bouchard

Covariate Selection Method in Propensity Score Model for the Quantile Treatment Effect Estimation 66
Takehiro Shoji, Jun Tsuchida, and Hiroshi Yadohisa

Are Attitudes Toward Immigration Changing in Europe? An Analysis Based on Latent Class IRT Models 67
Ewa Genge and Francesco Bartolucci

Visualization of IATA Regions in Air Transport Before and After the COVID-19 Pandemic 68
Tüzün Tolga İnan, Neslihan Gökmen İnan, Aylin Yaman Kocadağlı, and Ozan Kocadağlı

Political and Religion Attitudes in Greece: Behavioral Discourses 69
Georgia Panagiotidou and Theodore Chadjipadelis

Functional Data Representation with Merge Trees 70
Matteo Pegoraro and Piercesare Secchi

Elastic Regression for Irregularly Sampled Curves in $\mathbb{R}^d$ 71
Lisa Steyer, Almond Stöcker, and Sonja Greven

Misalignment of Spectral Data: Constrained Optimization in a Functional Data Analysis Framework 72
Francesca Di Salvo, Delia Francesca Chillura Martino, and Gabriella Chirco

Model Based Clustering of Functional Data with Mild Outliers 73
Cristina Anton and Iain Smith

Old and New Constraints in Model Based Clustering 74
Luis A. García-Escudero, Agustín Mayo-Iscar, Gianluca Morelli, and Marco Riani

Model Based Clustering and Outlier Detection with Missing Data 75
Cristina Tortora, Hung Tong, and Louis Tran

How to Mitigate the Effect of Outliers on Balancing Technique 76
Rasool Taban, Maria do Rosário Oliveira, and Claudia Nunes Philippart

Outliers Detection in Functional Data 77
Amovin-Assagba Martial, Gannaz Irène, and Jacques Julien

Robustified Elastic Net Estimator for Multinomial Regression 78
Fatma Sevinç Kurnaz and Peter Filzmoser

Optimized Symbolic Correspondence Analysis for Multi-valued Variables 79
Jorge Arce Garro and Oldemar Rodríguez Rojas

Symbolic t-SNE and UMAP Methods for Interval Type Variables. 80
Oldemar Rodríguez Rojas

Two-stage Principal Component Analysis on Interval-valued Data Using Patterned Covariance Structure 81
Anuradha Roy

Detection of the Biliary Atresia Using Deep Convolutional Neural Networks Based on Statistical Learning Weights via Optimal Similarity and Resampling Methods 82
Kuniyoshi Hayashi, Eri Hoshino, Mitsuyoshi Suzuki, Erika Nakanishi, Kotomi Sakai, and Masayuki Obatake

Variational Autoencoder with Gamma Mixture for Clustering Right-skewed Data 83
Jinwon Heo and Jangsun Baek

An Efficient Way to Identify Inliers via Inlier-memorization Effect of Deep Generative Models 84
Dongha Kim, Jaesung Hwang, and Yongdai Kim

Three-way Spectral Clustering 85
Cinzia Di Nuzzo and Salvatore Ingrassia

Fuzzy Clustering by Hyperbolic Smoothing 86
David Masís, Esteban Segura, Javier Trejos, and Adilson Xavier

Combining KDE and DBSCAN Clustering to Understand Road Traffic Accidents: the Case of Setúbal, Portugal 87
Pedro Nogueira, Marcelo Silva, Paulo Infante, Paulo Rebelo Manuel, Leonor Rego, Anabela Afonso, and Gonçalo Jacinto

Similarity Forest for Time Series Classification 88
Tomasz Górecki, Maciej Łuczak, and Paweł Piasecki

Uncovering Regions of Maximum Dissimilarity on Random Process Data 89
Miguel de Carvalho and Gabriel Martos Venturini

Franz Liszt's Transcendental Études: an Evolutionary Analysis by Machine Learning 90
Matteo Farnè

Quantile-distribution Functions and Their Use for Classification 91
Edoardo Redivo, Cinzia Viroli, and Alessio Farcomeni

Analysis of Gini Splitting Criterion and Comparison with Maximum Likelihood Rule 92
Amirah S. Alharthi and Charles C. Taylor

Envelope-based Support Vector Machine Classifier 93
Alya Alzahrani and Andreas Artemiou

A Moment-free Measure of Multivariate Skewness 94
Andrzej Sokolowski and Malgorzata Markowska

The Weighted RV Coefficient: Exact Moments by Invariant Orthogonal Integration 95
François Bavaud

Testing Equality of Multivariate Coefficients of Variation 96
Marc Ditzhaus and Łukasz Smaga

A New Regression Model for the Analysis of Microbiome Data 97
Roberto Ascari and Sonia Migliorati

The Death Process in Italy Before and During the Covid-19 Pandemic: a Functional Compositional Approach 98
Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli, and Piercesare Secchi

Sampling Design for Uncovering Natural Laws in Compositional Data . . . 99
Lan Liang, Glòria Mateu-Figueras, and Jan Graffelman

Penalized Model-based Functional Clustering: a Regularization Approach via Shrinkage Methods 100
Nicola Pronello, Rosaria Ignaccolo, Luigi Ippoliti, and Sara Fontanella

Clustering in FDA Mixing the Epigraph and the Hypograph Indexes with Machine Learning Algorithms 101
Belén Pulido, Alba M. Franco-Pereira, and Rosa E. Lillo

Localization Processes for Functional Data Classification 102
Antonio Elías, Raúl Jiménez, and Joseph E. Yukich

A New Functional Data Clustering Technique Based on Spectral Clustering and Downsampling 103
Maryam Al Alawi, Surajit Ray, and Mayetri Gupta

Parsimonious Mixtures of Seemingly Unrelated Contaminated Normal Regression Models 104
Gabriele Perrone and Gabriele Soffritti

Monitoring Hyperparameter Choice for Robust Cluster Weighted Model 105
Andrea Cappozzo, Luis A. García-Escudero, Francesca Greselin, and Agustín Mayo-Iscar

Latent Block Regression Model 106
Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif

Towards a Bi-stochastic Matrix Approximation of k-means and Some Variants 107
Lazhar Labiod and Mohamed Nadif

Clustering Brain Connectomes Through a Density-peak Approach 108
Riccardo Giubilei

New Metrics for Classifying Phylogenetic Trees Using k-means and the Symmetric Difference Metric 109
Nadia Tahiri and Aleksandr Koshkarov

Alternating Optimization Framework for Sparse Simultaneous Component Analysis Based on Data Integration 110
Rosember Guerra-Urzola, Juan C. Vera, Katrijn Van Deun, and Klaas Sijtsma

Joint Sparse Principal Component Analysis 111
Katrijn Van Deun

Joint Sparse Principal Component Analysis: a Simulation Study 112
Tra Le and Katrijn Van Deun

Copula-based Non-metric Unfolding on Augmented Data Matrix 113
Marta Nai Ruscone and Antonio D'Ambrosio

Emotion Classification Based on Single Electrode Brain Data: Applications for Assistive Technology 114
Duarte Rodrigues, Luis Paulo Reis, and Brígida Mónica Faria

On the Role of Data, Statistics and Decisions in a Pandemic 115
Ursula Garczarek, Beate Jahn, Sarah Friedrich, Joachim Behnke, Joachim Engel, Ralf Münnich, Markus Pauly, Adalbert Wilhelm, Olaf Wolkenhauer, Markus Zwick, Uwe Sieber, and Tim Friede

A Deep Learning Analytics to Detect Dental Caries 116
Taerim Lee

Identification of Shared Genetic Loci Between Psychiatric Disorders and Telomere Length and Evaluation of Their Role as Potential Drug Targets 117
Claudia Pisanu, Anna Meloni, and Alessio Squassina

Estimating Optimal Decision Trees for Treatment Assignment with $k > 2$ **Treatment Alternatives: a Classification Problem with a Unit- and Class- dependent Misclassification Cost** 118
Iven Van Mechelen and Aniek Sies

ExactTree: an R-package for Globally Optimal Decision Trees 119
Elise Dusseldorp, Juan Claramunt Gonzales, Jacqueline Meulman, Samil Uysal, and Bart Jan van Os

Optimal Random Projection Trees Ensemble 120
Nosheen Faiz, Adi Lausen, Metodi Metodiev, Zardad Khan, and Berthold Lausen

Born-again and Bayesian Approaches for Improving the Performance of Decision Trees 121
Marjolein Fokkema

Evolution of Media Coverage on Climate Change and Environmental Awareness: an Analysis of Tweets from UK and US Newspapers 122
Gianpaolo Zammarchi, Maurizio Romano, and Claudio Conversano

Improving Classification of Documents by Semi-supervised Clustering in a Semantic Space 123
Jasminka Dobša and Henk A.L. Kiers

Is It Hate or Criticism? An Exploratory Approach to Negative Comments on YouTube 124
Manuela Schmidt

A Time-varying Text Based Ideal Point Model to Infer Partisanship in the U.S. Senate 125
Sourav Adhikari, Bettina Grün, and Paul Hofmarcher

Oracle-LSTM: a Neural Network Approach to Mixed Frequency Time Series Prediction 126
Alessandro Bitetto and Paola Cerchiello

Time Series of Counts Under Censoring 127
Isabel Silva, Maria Eduarda Silva, Isabel Pereira, and Brendan McCabe

Multivariate Time Series Feature Extraction via Multilayer Networks . . . 128
Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, and Fernando Silva

On the Use of the Choquet Fuzzy Integral to Aggregate Predictions of Time Series: an Application to Economic (and Other Types of) Data 129
Diogo Alves, José Matos, and Sandra Silva

Some Biplot Alternatives 130
Patrick Groenen

Biplots in Dimension Reduction and Clustering 131
Alfonso Iodice D'Enza, Angelos Markos, and Michel van de Velden

Biplots: a Sophisticated Multivariate Approach or a User-Friendly Technique? 132
Manuel Rui Alves

PLS-based Principal Balances for Regression and Classification with High-dimensional Compositional Data 133
Viktorie Nesrstová, Ines Wilms, Karel Hron, Josep A. Martín-Fernández, Peter Filzmoser, and Javier Palarea-Albaladejo

Clustering Count Data Using Compositional Methods 134
Marc Comas-Cufí, Josep A. Martín-Fernández, Glària Mateu-Figueras, and Javier Palarea-Albaladejo

Urban Development Paths in Poland: Multidimensional Perspective 135
Barbara Batóg and Jacek Batóg

Analyzing the Evolution of EU Countries and Indicators of Europe 2020 Agenda 136
Adelaide Figueiredo and Fernanda Figueiredo

Google Trends as a Macroeconomic Predictor: Behind the Scenes 137
Eduardo Andre Costa and Maria Eduarda Silva

COVID-19 Pandemic: a Methodological Model for the Analysis of Government Preventing Measures and Health Data Records 138
Theodore Chadjipadelis and Sofia Magopoulou

Detecting Fabricated Interviews Using the Hamming Distance 139
Joerg Blasius

Digital Development and Internet Use in the European Union Countries . 140
Fernanda Figueiredo and Adelaide Figueiredo

Probabilistic Clustering with Local Alignement of Italian COVID-19 Death Curves 141
Marzia A. Cremona, Tobia Boschi, and Francesca Chiaromonte

Model Free Predictive Inference for Functional Kriging Techniques Based on Conformal Prediction 142
Andrea Diana, Elvira Romano, and Jorge Mateu

Density Modelling via Functional Data Analysis 143
Stefano A. Gattone and Tonio Di Battista

On Parsimonious Modelling via Matrix-variate t Mixtures 144
Salvatore D. Tomarchio

Four Skewed Tensor Variate Distributions 145
Michael P.B. Gallaugher, Peter A. Tait, and Paul D. McNicholas

A Family of Skewed Power Exponential Mixture Models for Clustering and Classification 146
Utkarsh J. Dang, Michael P. B. Gallaugher, Ryan P. Browne, and Paul D. McNicholas

Generating Collective Counterfactual Explanations in Score-based Classification via Mathematical Optimization 147
Jasone Ramírez-Ayerbe, Emilio Carrizosa, and Dolores Romero Morales

Spherical Separation in Machine Learning 148
Matteo Avolio, Annabella Astorino, and Antonio Fuduli

Model Extraction Based on Counterfactual Explanations 149
Cecilia Salvatore and Veronica Piccialli

Isolation Forests for Symbolic Data as a Tool for Outlier Mining 150
Andrzej Dudek and Marcin Pelka

Symbolic Clustering Methods Applied to Interval Estimates of Production Cost Quantiles 151
Dominique Desbois

Fisher Discriminant Analysis for Interval Data 152
Diogo Pinheiro, Maria do Rosário Oliveira, Igor Kravchenko, and Lina Oliveira

Stability of Mixed-type Cluster Partitions for Determination of the Number of Clusters 153
Rabea Aschenbruck, Gero Szepannek, and Adalbert Wilhelm

Multinomial Multilevel Models with Discrete Random Effects: a Multivariate Clustering Tool 154
Chiara Masci, Francesca Ieva, and Anna Maria Paganoni

PD-clustering for Mixed Data Type 155
Francesco Palumbo and Cristina Tortora

Anomaly Detection-based Under-sampling for Imbalanced Classification Problems 156
You-Jin Park, Chun-Yang Peng, Rong Pan, and Douglas C. Montgomery

Continuous Adaptation to Distribution Drifts Through Continual Learning in Manufacturing 157
Henrique Siqueira and Onay Urfalioglu

Detecting Anomalies with TADGAN: a Case Study 158
Inês Oliveira e Silva, Carlos Soares, Arlete Rodrigues, and Pedro Bastardo

A Trivariate Geometric Classification of Decision Boundaries for Mixtures of Regressions 159
Filippo Antonazzo and Salvatore Ingrassia

Clustering Rainfall by Simulated Annealing for Histogram Symbolic Data 160
Alejandro Chacón and Javier Trejos

Statistical Assessment of Youth Inclusion in the National Labour Markets 161
Beata Bal-Domańska

Barriers to Industry Digitization in Poland from the Perspective of High and Medium-high Technology Sector Enterprises 162
Elżbieta Sobczak, Marcin Pelka, and Karolina Pokorska

Kernel Smoothing-based Probability Contours for Tumour Segmentation 163
Wenhui Zhang and Surajit Ray

Parameter Estimation for Mixtures of Linear Mixed Models: the EM, CEM and SEM Algorithms 164
Luísa Novais and Susana Faria

Genomic Prediction Using Machine Learning Methods: Performance Comparison on Synthetic and Empirical Data 165
Vanda Lourenço, Joseph Ogutu, Rui Rodrigues, and Hans-Peter Piepho

Pooled Mean and Confidence Interval Estimation Combining Different Sets of Summary Statistics 166
Flora Ferreira, José Soares, Fernanda Sousa, Filipe Magalhães, Isabel Ribeiro, and Pedro Pacheco

Prediction of Diabetes via Bayesian Network Classifier from Exposure to Environmental Polluting Chemicals Data 167
Rosy Oh, Hong Kyu Lee, Youngmi Kim Pak, and Man-Suk Oh

Model Performance Metrics for Sample Selection Bias Correction by Pseudo Weighting 168
An-Chiao Liu, Ton de Waal, Katrijn Van Deun, and Sander Scholtus

Some Factors that Influence the Nature of Road Traffic Accidents 169
Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Leonor Rego, Vitor Nogueira, Paulo Quaresma, José Saias, Daniel Santos, Pedro Nogueira, Marcelo Silva, Rosalina Pisco Costa, Patrícia Gois, and Paulo Rebelo Manuel

Comparing Variable Selection Methods for High-dimensional Compositional Data in a Discriminant Analysis Context 170
Pepus Daunis-i-Estadella, Glòria Mateu-Figueras, Viktorie Nesrstová, Karel Hron, and Josep A. Martín-Fernández

Cluster Analysis and Genetic Risk Score in Age-related Macular Degeneration - the Coimbra Eye Study 171
Rita Coimbra

Sensor System for Standardizing Articulation Patterns According to Korean Phonemes 172
Seong Tak Woo and Da Hee Oh

Categorical Data Visualization and the Cressie-Read Divergence Statistic 173
Eric J. Beh and Rosaria Lombardo

Biplots Based on Latent Variable Models in the Analysis of Ecological Communities 174
Jenni Niku and Sara Taskinen

Biplot Representation of Partial Least Squares Regression for Binary Responses 175
Laura Vicente-Gonzalez and Jose Luis Vicente-Villardon

Generalized Spatio-temporal Regression with PDE Penalization 176
Eleonora Arnone, Elia Cunial, and Laura M. Sangalli

Impact Point Selection in Semiparametric Bifunctional Models 177
Silvia Novo, Germán Aneiros, and Philippe Vieu

Latent Function-on-scalar Regression Models for Observed Sequences of Correlated Binary Data: a Restricted Likelihood Approach 178
Fatemeh Asgari and Valeria Vitelli

pcTVI: Parallel MDP Solver Using a Decomposition into Independent Chains 179
Jaël Champagne Gareau, Éric Beaudry, and Vladimir Makarenkov

Classification of Viral Pneumonia Images via Multiple Instance Learning 180
Antonio Fuduli, Matteo Avolio, Eugenio Vocaturo, and Ester Zumpano

Nonlinear Approaches for Multiple Instance Learning 181
Annabella Astorino, Matteo Avolio, and Antonio Fuduli

An Online Minorization-Maximization Algorithm 182
Hien Duy Nguyen, Florence Forbes, Gersende Fort, and Olivier Cappé

Frugal Gaussian Clustering of Huge Ombalanced Datasets Through a Bin-marginal Approach 183
Filippo Antonazzo, Christophe Biernacki, and Christine Keribin

Reinforced EM Algorithm Through Clever Initialization for Clustering with Gaussian Mixture Models 184
Joshua Tobin, Chin Pang Ho, and Mimi Zhang

Exact Computation of the Angular Halfspace Depth 185
Stanislav Nagy and Rainer Dyckerhoff

Reconstruction of Atomic Measure Based on its Simplicial Depth 186
Petra Laketa and Stanislav Nagy

Robustness Aspects of Optimized Centroids 187
Jan Kalina and Patrik Janáček

Analysis of the Changes in the Polish Traditional Drugstores Market During COVID-19 188
Marcin Pelka, Antonio Irpino, and Michal Swachta

Logistic Regression Models for Aggregated Data 189
Thomas Whitaker, Boris Beranger, and Scott Sisson

Nonparametric Regressions for Distributional Data 190
Albert Meco, Javier Arroyo, and Antonio Irpino

Hotspot Cluster Detection Based on Spatial Hierarchical Structure and its Software 191
Fumio Ishioka, Shoji Kajinishi, and Koji Kurihara

Group Lasso Penalty for Spatially Clustered Coefficient Regression 192
Toshiki Sakai, Jun Tsuchida, and Hiroshi Yadohisa

Visualization of the Number of New Positives for COVID-19 in Japan . . . 193
Yoshiro Yamamoto, Sanetoshi Yamada, Mayumi Tanahashi, and Tadashi Imanishi

Two Simple but Efficient Elgorithms to Recognize Robinson Dissimilarities 194
Mikhaël Carmona, Guyslain Naves, Victor Chepoi, and Pascal Préa

Clustering with Missing Data: Which Imputation Model for Which Cluster Analysis Method? 195
Vincent Audigier, Ndèye Niang, and Matthieu Resche-Rigon

Hierarchies and Weak-hierarchies as Interval Convexities 196
Patrice Bertrand and Jean Diatta

A Rule-based Approach to Scoring Systems 197
Michael Rapp, Johannes Fürnkranz, and Eyke Hüllermeier

On Explaining Model Change Based on Feature Importance 198
Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier

Interpretable Multi-class Trees for Travel Choice Mode Analysis 199
Christian Riccio, Andrea Papola, Michele Staiano, and Roberta Siciliano

Identification of Driver Genes in Glioblastoma via Regularized Classification 200
Marta Belchior Lopes and Susana Vinga

Outlier Detection: a Procedure to Capture Atypical Groups of Observations 201
Ana Helena Tavares, Vera Afreixo, and Paula Brito

Bayesian Classification and Non-Bayesian Label Estimation via EM Algorithm to Identify Differential Expression in Omics Data: a Comparative Study 202
Marilia Antunes and Lisete Sousa

Sequence-aware Item Recommendations for Multiply Repeated User-item Interactions 203
Juan Pablo Equihua, Maged Ali, Henrik Nordmark, and Berthold Lausen

High-dimensional Linear Regression Estimation 204
Mauro Iannuzzi and Matteo Farnè

Experimental Study of Similarity Measures for Clustering Uncertain Time Series 205
Michael Dinzinger, Michael Franklin Mbouopda, and Engelbert Mephu Nguifo

Assessing the Status of Two Data-limited Skates Landed in Portuguese Ports Using an Empirical Catch Rule 206
Erick Chatalov, Ivone Figueiredo, Lisete Sousa, and Bárbara Pereira

Machine Learning Approach to Identify Factors that Influence Accident Severity 207
Daniel Santos, Vitor Nogueira, José Saias, Paulo Quaresma, Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Leonor Rego, Pedro Nogueira, Marcelo Silva, Rosalina Pisco Costa, Patrícia Gois, and Paulo Rebelo Manuel

Trade and Bank Credit of Portuguese SMEs: a Panel Data Application . . 208
Carla Henriques, Pedro Pinto, and Carolina Cardoso

Hausdorff Distance: a Powerful Tool for Matching Households and Individuals in Historical Censuses 209
Thais Pacheco Menezes, Michael Fop, and Thomas Brendan Murphy

Model-based Tri-clustering 210
Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif

Analyzing the Effects of Deviations from Normality on the Latent Growth Curve Models Goodness-of-fit 211
Catarina Marques, Maria de Fátima Salgueiro, and Paula C.R. Vicente

Transformation Mixture Modeling for Skewed Data Groups with Heavy Tails 212
Yana Melnykov, Xuwen Zhu, and Volodymyr Melnykov

A Simulation Study on Variable Selection in Mixture Regression Models . 213
Susana Faria

Hybrid Forecasting Combinations by Feature Based Metalearning 214
Moises Santo, Andre C.P.L.F. de Carvalho, and Carlos Soares

On the Measurement of Household Subjective Poverty: Concepts and Application 215
Aleksandra Łuczak and Sławomir Kalinowski

Socio-economic Classification of Territorial Units: Extreme Value Theory-based Methods as Support for the Construction of a Synthetic Index 216
Aleksandra Łuczak and Małgorzata Just

Depth-based Two-sample Testing 217
Felix Gnettner, Claudia Kirch, and Alicia Nieto-Reyes

The Control of False Discovery Rate for Functional Data 218
Niels Lundtorp Olsen, Alessia Pini, and Simone Vantini

Functional Random Forest for Biomedical Signals Classification and Interpretative Tools 219
Fabrizio Maturo and Rosanna Verde

Correlation-based Iterative Clustering Methods for Time Course Data . . 220
Michelle Carey, Shuang Wu, Guojun Gan, and Hulin Wu

Depth-based Classifiers for Partially Observed Functional Data 221
Antonio Elías, Raúl Jiménez, Anna Maria Paganoni, and Laura M. Sangalli

Using Clustering and Machine Learning Methods to Provide Intelligent Grocery Shopping Recommendations 222
Nail Chabane, Mohamed Achraf Bouaoune, Reda Amir Sofiane Tighilt, Bogdan Mazoure, Nadia Tahiri, and Vladimir Makarenkov

Typology of Motivation Factors for Employees in the Banking Sector: Multivariate Data Analysis 223
Áurea Sousa, Osvaldo Silva, M. Graça Batista, Sara Cabral, and Helena Bacelar-Nicolau

Industry Sector Detection in Legal Articles Using Transformer-based Deep Learning 224
Hui Yang, Stella Hadjiantoni, Yunfei Long, Rūta Petraitytė, and Berthold Lausen

User Segmentation Based on Online Behavioural Data via Ensemble Predictions and Clustering 225
Stella Hadjiantoni, Hui Yang,Yunfei Long, Ruta Petraityte, and Berthold Lausen

Attitudes Toward Statistics in the 3rd Cycle of Basic Education in Portugal 226
Adelaide Freitas, Ana Julieta Morais, Pedro Sá Couto, and Anabela Rocha

Predictors of Quantitative Skills in Degree Schemes at University 227
Alex Partner, Adi Lausen, Alexei Vernitski, Chris Saker, and Berthold Lausen

Using Excel and R for Teaching Statistics and Data Analysis 228
W.H. Moolman

Students' Assessment Through a IRT and Archetypal Analysis Joint Strategy 229
Lucio Palazzo and Francesco Palumbo

Kernel-based Hierarchical Structural Component Models for Pathway Analysis 230
Suhyun Hwangbo, Sungyoung Lee, Seungyeoun Lee, Heungsun Hwang, Inyoung Kim, and Taesung Park

Bayesian Inference for the Generation Interval of COVID-19 in Busan, Korea 231
Jayoeng Paek, Ilsu Choi, Kyeongah Nah, and Yongkuk Kim

Fitting an Accelerated Failure Time Model with Time-dependent Covariates via Nonparametric Gaussian Scale Mixtures 232
Ju-Young Park, Byungtae Seo, and Sangwook Kang

Comparison of Survival Prediction Models for Pancreatic Cancer: Cox Model vs. Machine Learning Models 233
Hyunsuk Kim, Taesung Park, and Seungyeoun Lee

Clustering High-dimensional Microbiome Data 234
Sanjeena Dang (Subedi) and Wangshu Tu

Clustering Adolescent Female Physical Activity Levels with an Infinite Mixture Model on Random Effects 235
Amy LaLonde, Tanzy Love, Deborah Rohm Young, and Tongtong Wu

Modeling Three-way RNA Sequencing Data Using Data Transformations and Matrix-variate Gaussian Mixture Models 236
Theresa Scharl and Bettina Grün

Some Issues in Robust Clustering 237
Christian Hennig

Assessing Common Principal Directions 238
David Rodríguez Vítores and Carlos Matrán Bea

Robustness and Initialization Issues in Subspace Clustering 239
Luis A. García-Escudero and Agustín Mayo-Iscar

A Likelihood Ratio Test for Choosing Input Parameters in Robust Model Based Clustering 240
Luis A. García-Escudero, Agustín Mayo-Iscar, Gianluca Morelli, and Marco Riani

Data Clustering and Representation Learning Based on Networked Data 241
Lazhar Labiod and Mohamed Nadif

Exploratory Graph Analysis for Configural Invariance Assessment of a Test 242
Alex Cucco, Lara Fontanella, Sara Fontanella, and Nicola Pronello

An Extension of Edge Reduction for Large Networks 243
Pedro Campos

Patterns of Cooperation for Polish Authors of Research Publications in Economics, Business and Medicine Areas 244
Urszula Cieraszewska, Paweł Lula, Magdalena Talaga, and Marcela Zembura

A Topological Clustering of Individuals 245
Rafik Abdesselam

Modeling a Most Specific Generalization in Domain Taxonomies 246
Zhirayr Hayrapetyan, Boris Mirkin, Susana Nascimento, trevor Fenner, and Dmitry Frotov

A Proposal for Formalization and Definition of Anomalies in Dynamical Systems 247
Jan Michael Spoor, Jens Weber, and Jivka Ovtcharova

Unsupervised Classification of Categorical Time Series Through Innovative Distances 248
Ángel López-Oriona, José A. Vilar, and Pierpaolo D'Urso

Detecting Differences in Italian Regional Health Services During Two Covid-19 Waves 249
Lucio Palazzo and Riccardo Ievoli

The Clustering Performance of a Weighted Combined Distance Between Time Series 250
Margarida G. M. S. Cardoso, Ana Alexandra Martins, and João Lagarto

Dimensionality Reduction and Multivariate Time Series Classification . . . 251
Veronne Yepmo, Angeline Plaud, and Engelbert Mephu Nguifo

A Review on Official Survey Item Classification for Mixed-Mode Effects Adjustment 252
Afshin Ashofteh and Pedro Campos

Adaptive Fuzzy Systems in Economics and Finance: Evaluating Interval Forecasts of High-frequency Data 253
Rosangela Ballini

The Usefulness of Selected Machine Learning Methods for Estimating Missing Data to Supplement Databases Used for Corporate Bankruptcy Prediction 254
Barbara Pawelek and Jozef Pociecha

Registration of 24-hour Accelerometric Rest-activity Profiles and its Application to Human Chronotypes 255
Erin I. McDonnell, Vadim Zipunnikov, Jennifer A. Schrack, Jeff Goldsmith, and Julia Wrobel

Functional Data from Wearable Devices: a Review 256
Nihan Acar-Denizli and Pedro Delicado

A Wavelet-mixed Effect Landmark Model for the Effect of Potassium and Biomarkers Profiles on Survival in Heart Failure Patients 257
Caterina Gregorio, Giulia Barbati, and Francesca Ieva

True Sparsity Approaches in Classification via Conic Optimization 258
Immanuel M. Bomze and Bo Peng

Creating Homogeneous Sectors: Criteria and Applications of Sectorization 259
Cristina Lopes, Maria Margarida Lima, Elif Göksu Öztürk, Ana Maria Rodrigues, Ana Catarina Nunes, Cristina Oliveira, José Soeiro Ferreira, and Pedro Filipe Rocha

MARGOT: a Maximum MARGin Optimal Classification Tree 260
Federico D'Onofrio, Marta Monaci, Giorgio Grani, and Laura Palagi

Multivariate Mapping of Soil Organic Carbon and Nitrogen 261
Stephan van der Westhuizen, David P. Hofmeyr, and Gerard B.M. Heuvelink

Spatial Configuration of Fire Stations in Portugal 262
Regina Bispo, Clara Yokochi, Francisca G. Vieira, Nádia Bachir, Pedro Espadinha-Cruz, José Pedro Lopes, Alexandre Penha, Marta Belchior Lopes, Filipe J. Marques, João Paulo Rodrigues, and António Grilo

Spatio-temporal Variability of Distribution and Abundance of Sardine in Portuguese Continental Coast: Environmental Effects 263
Daniela Silva, Raquel Menezes, Ana Moreno, Ana Teles-Machado, and Susana Garrido

Time Resolved Feature Importance of a Biopharmaceutical Purification Process Using Permutation Based Methods 264
Matthias Medl, Theresa Scharl, Astrid Dürauer, and Friedrich Leisch

Off-target Predictions in CRISPR-Cas9 Gene Editing Using Machine Learning 265
Ali Mertcan Kose and Ozan Kocadağlı

Comparison of k-mer and Alignment-based Pre-processing Approaches for Machine Learning Based Functional Annotation with 16S rRNA Data 266
Rafal Kulakowski, Adi Lausen, Etienne Low-Decarie, and Berthold Lausen

An Ultrametric Model for Clustering and Dimensionality Reduction 267
Giorgia Zaccaria

Combining Latent Class Analysis and Multiple Correspondence Analysis 268
Alice Barth

Simultaneous Factorial Reduction and Clustering for Three-mode Data Sets: a Comparison 269
Prosper Ablordeppey, Adelaide Freitas, Maurizio Vichi, and Giorgia Zaccaria

Clustering Intensive Longitudinal Data Through Mixture Multilevel Vector-autoregressive Modeling 270
Anja Ernst, Marieke Timmerman, Feng Ji, Bertus Jeronimus, and Casper Albers

Mispecification Tests for Hidden Markov Models Based on a New Class of Finite Mixture Models 271
Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni

Natural Cubic Smoothing Splines for Latent Class Identification in Longitudinal Growth Trajectories 272
Katerina M. Marcoulides and Laura Trinchera

Supervised Classification via Neural Networks for Replicated Point Patterns 273
Kateřina Pawlasová, Iva Karafiátová, and Jiří Dvořák

Reliability Assessment of Ancient Stone Arch Bridge Applying ANN models, Case Study: Leça Railway Bridge 274
Edward A. Baron, Ana Margarida Bento, José Campos e Matos, Rui Calçada, and Kenneth Gavin

Application of Artificial Intelligence (AI) in Flood Risk Forecasting 275
Minh Quang Tran, Ana Margarida Bento, Elisabete Teixeira, Hélder Sousa, and José Campos e Matos

Logistic Regression with Sparse Common and Distinctive Covariates . . . 276
Soogeun Park, Eva Ceulemans, and Katrijn Van Deun

Accuracy Measures for Binary Classification Based on Quantitative Group Tests 277
Rui Santos, João Paulo Martins, and Miguel Felgueiras

Exploiting Pareto Density Estimation for Nonparametric Naïve Bayes Classifiers 278
Quirin Stier and Michael C. Thrun

Author Index 279

Part I
Presidential Address

Perturb and Conquer: how Classification Can Benefit from Data Perturbation

Angela Montanari

Data perturbation has a longstanding tradition in statistics. The bootstrap method and random forests are evergreen examples in this stream. The focus, in this talk, will be on the use of data perturbation for classification purposes. In particular, we will focus on perturbations obtained by random projections [1] and we will address issues ranging from variable selection [2] to imbalanced classes [3], from data shift to semi-supervised learning.

The content of the talk is a joint work with Laura Anderlucci, Department of Statistical Sciences University of Bologna.

Keywords: random projections, data shift, imbalanced classes

References

1. Cannings, T. I., Samworth, R. J.: Random-projection ensemble classification. J. R. Stat. Soc. B **79**, 959–1035 (2017)
2. Fortunato, F., Anderlucci, L., Montanari, A.: One-class classification with application to forensic analysis. J. R. Stat. Soc. C **69**, 1227–1249 (2020)
3. Falcone, R., Anderlucci, L., Montanari, A.: Matrix sketching for supervised classification with imbalanced classes. Data Mining and Knowledge Discovery **36**, 174–208 (2022)

Angela Montanari
Department of Statistical Sciences - University of Bologna, via Belle Arti 41,
e-mail: angela.montanari@unibo.it

Part II
IFCS Research Medal Lecture

An Introduction to S-concordance and S-discordance

Edwin Diday

In the 17th century, Galileo Galilei gave a duty to humanity by saying that we have to measure everything that is measurable and "make measurable" everything which is not. Here, our aim is to "make measurable" the notions of "concordance" and "discordance" usually used between behaviours, events, ideas, results, etc. For example, we can measure the "concordance" between a country and the European countries for some given variables related to corona. Also, the "concordance" of a solar star with a collection of solar stars having a planet, the "concordance" of a stock and a portfolio of stocks in their behaviour during a given period, a new species with a family of species, etc. The Kendall "concordance" and "discordance" are used to compare ordinal variables, that is why "s-concordance" and "s-discordance" are used here ("s" for "symbolic" as symbolic data are used). The "s-concordance" or "s-discordance" definition between a class c and a collection P of given classes for a given value x (or vector of values), needs two basic densities. Roughly said, the first is $f_c(x)$ which expresses the proportion of the x value of a descriptive variable (which can be multidimensional) inside the class c and the second is $g_x(c)$ which expresses the proportion of classes of P which have a proportion of the x value equal or close to $f_c(x)$. Then, the "s-concordance" (resp. s-discordance) denoted $S_{conc}(c,P,x)$ (resp. $S_{disc}(c,P,x)$) satisfies natural axioms. From the given data, specific families of concordance can be induced by using the copulas obtained from the joint distribution function of the random variables which densities are f_c and g_x. S-concordance and s-discordance differ from similarities and dissimilarities by their meaning but also by their axiomatic definition. The s-discordance has as a specific case the Tf-Idf. S-concordance and s-discordance have numerous impacts in data science. These impacts are illustrated in the case of the k-means, dynamical clustering, hierarchical or pyramidal clustering, mixture decomposition, Latent Dirichlet Allocation, statistical inference and likelihood. A potential application in genomics is presented.

Keywords: symbolic data analysis, s-concordance, s-discordance, copulas

References

1. Diday, E.: Explanatory tools for Machine Learning in the Symbolic Data Analysis framework. In: Diday, E., Guan, R., Saporta, G., Wang, H. (eds.) Advances in Data Science. ISTE-Wiley (2020)
2. Nelsen, R.B. An Introduction to Copulas. Lecture Notes in Statistics. Springer (1998)

Edwin Diday
CEREMADE, University Paris-Dauphine, Paris, France, e-mail: diday8@gmail.com

Part III
Tutorials

Analysis of Data Streams

João Gama

The challenge of deriving insights from the Internet of Things (IoT) has been recognized as one of the most exciting and key opportunities for both academia and industry. Advanced analysis of big data streams from sensors and devices is bound to become a key area of data mining research as the number of applications requiring such processing increases. Dealing with the evolution over time of such data streams, i.e., with concepts that drift or change completely, is one of the core issues in IoT stream mining. This tutorial is a gentle introduction to mining IoT big data streams.

Content: IoT Fundamentals and Stream Mining Algorithms; IoT Stream mining setting; Clustering; Classification and Regression; Concept drift and Frequent Pattern mining.

J. Gama
FEP-University of Porto, INESC TEC, e-mail: jgama@fep.up.pt

Categorical Data Analysis and Visualization

Rosaria Lombardo and Eric J. Beh

The impact of the internet, social media and smart devices means that people are becoming increasingly literate with use of these technologies and it has changed how we engage with others on a professional and personal level. For the analyst, the capacity to adapt to such changes has impacted upon the tools designed for analyzing "big data", i.e. huge amount of numerical and categorical data. One of the most important tools is that of "visualization".
With focus on categorical data, this tutorial, after briefly introducing association indices, models and methods, will outline some cutting-edge visualization tools and techniques.

Content: A Quick Historical Overview of the Visualization of Categorical Data; The Contingency Table and the Chi-Squared Statistic; Measures of Symmetric Association for I x J Contingency Tables; Correspondence Analysis (symmetrical, non-symmetrical, ordinal) and Multiple and Multi-way Correspondence Analysis.

Rosaria Lombardo
Dipartimento di Economia, Università degli Studi della Campania Luigi Vanvitelli
e-mail: rosaria.lombardo@unicampania.it

Erich Beh
School of Information and Physical Science, University of Newcastle, Australia

Part IV
Keynote Lectures

A Showcase of New Methods for High Dimensional Data Viewing with Linear Projections and Sections

Dianne Cook

In the last few years there have been several huge strides in new methods available for exploring high-dimensional data using tours. Tours is the collective term for visualisations built on linear projections. Tours have two key elements: the *path* that generates a sequence, and the *display* to make of the low-dimensional projection. There are numerous path algorithms available (and implemented in the tourr [1] R package), including the old (grand, guided, little, local, manual), and the new (slice [2], sage [3]). This talk will show off these new tools and how they can be used for several contemporary problems, including understanding nonlinear dimension reductions (e.g. t-SNE) using the liminal [4] R package, and explainable artificial intelligence (XAI) using the cheem [5] R package. Step into the fascinating world of high-dimensions with me.

This most recent methodology is joint with primarily Ursula Laa, German Valencia, Stuart Lee and Nicholas Spyrison.

Keywords: statistical graphics, exploratory data analysis, tours, XAI, t-SNE, R

References

1. H. Wickham, D. **Cook**, H. Hofmann, and A. Buja. tourr: An R package for exploring multivariate data with projections. *Journal of Statistical Software*, 40(2):1–18, 2011. URL http://www.jstatsoft.org/v40/i02/.
2. U. Laa, D. Cook, and G. Valencia. A slice tour for finding hollowness in high-dimensional data. *Journal of Computational and Graphical Statistics*, 29(3):681–687, 2020. URL https://doi.org/10.1080/10618600.2020.1777140.
3. U. Laa, D. Cook, and S. Lee. Burning sage: Reversing the curse of dimensionality in the visualization of high-dimensional data. *Journal of Computational and Graphical Statistics*, 31(1):40–49, 2022. URL https://doi.org/10.1080/10618600.2021.1963264.
4. S. Lee. liminal: Multivariate data visualization with tours and embeddings, 2021. URL https://CRAN.R-project.org/package=liminal. R package version 0.1.2.
5. N. Spyrison cheem: Interactively Explore the Support of Local Explanations with the Radial Tour, 2022. URL https://CRAN.R-project.org/package=cheem. R package version 0.2.0.

Dianne Cook
Monash University, Melbourne, Australia, e-mail: dicook@monash.edu

Statistical Learning with Dynamic Interaction Data for Public Health

Charles Bouveyron

This work focuses on the problem of statistical learning with dynamic relational data for two specific public health problems: the analysis of the COVID-19 publication network and the study of a large-scale pharmacovigilance data set. On the one hand, the Covid-19 epidemic presented a unique use case for researchers and institutions in the health field where the ability to monitor and synthesize scientific publications on a given theme has proven to be strategic. Indeed, with more than 5000 publications and pre-publications per month on the Covid-19 virus, it has proved essential for researchers and doctors to have tools capable of synthesizing publications on this subject by grouping them on the basis of the research themes they mobilize. On the other hand, pharmacovigilance is a central medical discipline aiming at monitoring and detecting public health events (adverse drug reactions) caused by medicines and vaccines. As the current expert detection of safety signals is unfortunately incomplete due to the workload it represents, we investigate here an automatized method of safety signal detection from ADR data. To address those problems, we proposed two generative models for clustering the nodes of a dynamic graph, accounting for the content of textual edges as well as their frequency, in the first case, and the co-clustering of dynamic count data, for pharmacovigilance. In both cases, the continuous time is handled by partitioning the considered time period, allowing the detection of temporal breaks in the signals.

Keywords: clustering, interaction data, dynamic data, generative model, EM algorithm

Acknowledgements This work has been supported by the French government, through the 3IA Côte dAzur Investment in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

1. C. Bouveyron, M. Corneli, P. Latouche and F. Rossi: The dynamic stochastic topic block model for dynamic networks with textual edges, Statistics and Computing, vol. 29, pp. 677–695 (2019).
2. G. Marchello, A. Fresse, M. Corneli and C. Bouveyron: Co-clustering of evolving count matrices in pharmacovigilance with the dynamic latent block model, Preprint HAL 03146769, Université Côte d'Azur (2021).

Charles Bouveyron
Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team, France
e-mail: charles.bouveyron@univ-cotedazur.fr.

Fast Minipatch Ensemble Strategies for Learning and Inference

Genevera I. Allen

Enormous quantities of data are collected in many industries and disciplines; this data holds the key to solving critical societal and scientific problems. Yet, fitting models to make discoveries from this huge data often poses both computational and statistical challenges. In this talk, we propose a new ensemble learning strategy primed for fast, distributed, and memory-efficient computation that also has many statistical advantages. Inspired by random forests, stability selection, and stochastic optimization, we propose to build ensembles based on tiny subsamples of both observations and features that we term minipatches. While minipatch learning can easily be applied to prediction tasks similarly to random forests, this talk focuses on using minipatch ensemble approaches in unconventional ways: We will highlight new minipatch learning methods for unsupervised learning, specifically clustering and structural graph learning, and distribution-free and model-agnostic inference for both predictions and important features. Through huge real data examples from neuroscience, genomics and biomedicine, we illustrate the computational and statistical advantages of our minipatch ensemble learning strategies.

Keywords: ensemble learning, consensus clustering, graphical model selection, conformal inference, feature importance inference

Genevera I. Allen
Rice University, Houston, TX USA; e-mail: gallen@rice.edu

Trends in Data Stream Mining

João Gama

Learning from data streams is a hot topic in machine learning and data mining. We describe our recent work on emerging issues related to learning from data streams. We discuss two quite different problems. The first use case is an application of data stream techniques to fraud detection. We propose an algorithm for the interconnected by-pass fraud problem. This real-world problem requires processing high-speed telecommunications data and providing fraud alarms in real-time. The proposed solution clearly illustrates the need for online data stream processing.

Hyper-parameter tunning is a popular topic in offline learning. Nevertheless, few algorithms have been presented for the online setting. We present one of the first algorithms for online hyper-parameter tuning for streaming data. We discuss the Self hyper-Parameter Tunning (SPT) algorithm, an optimization algorithm for online hyper-parameter tuning from non-stationary data streams. SPT works as a wrapper over any streaming algorithm and can be used for classification, regression, and recommendation.

Keywords: fraud detection, hyper-parameter tuning, learning from data streams

J. Gama
FEP-University of Porto, INESC TEC, e-mail: jgama@fep.up.pt

Part V
Invited Sessions

A Criterion for Selecting the Number of Time Series Clusters

Daniel Peña and Ruey S. Tsay

A new method is proposed to select the number of clusters in hierarchical clustering of a set of independent time series, including a test statistic for detecting existence of multiple cluster. The method focuses on the steps (height increments) of a dendrogram and uses simulation to generate a reference distribution of the step. The proposed test statistic employs an upper sample quantile of the dendrogram steps of the data and the reference distribution. The largest step of the dendrogram is then used to select the number of clusters. We provide theoretical justification for the proposed method and show that it works well in simulation and applications. The performance of the criterion is illustrated with different measures of similarity between the univariate time series features.

Keywords: dendrogram heights; hierarchical clustering; linear time series models

Daniel Peña
Department of Statistics, Universidad Carlos III de Madrid, e-mail: daniel.pena@uc3m.es

Ruey S. Tsay
Booth School of Business, University of Chicago, e-mail: Ruey.Tsay@chicagobooth.edu

Clusters Based on Prediction Accuracy of Global Forecasting Models

Pablo Montero-Manso, Ángel López-Oriona, and José A. Vilar

In the context of model-based time series clustering, the quality of a given clustering relies on the predictive accuracy of the models that generated the cluster. Traditionally, a model is fitted to each time series and then a dissimilarity matrix is generated from distances between models. This approach has a few limitations. Time series are notoriously difficult to fit, exhibiting problems such temporal dependency, low sample size and nonstationarity. Moreover, when clustering a large number time series, we often have to rely on automatic fitting procedures without human supervision. These problems lead to models with poor predictive accuracy. However, a new paradigm has emerged in time series prediction, the so-called cross-learning or global model approach. A group of time series is pooled together and fitted with a single model, called a 'global' model, and a single predictive function is then used for all of the time series in the group. Global models are showing superior accuracy compared with traditional single-series (or 'local') models in a vast number of applications. Global models can be used for clustering, by finding a grouping that maximizes the predictive accuracy of the global models fitted to each group. This approach has an important side-effect, it introduces the conceptof predictive accuracy as a measure of a cluster quality: given a model class and a dataset,the quality of a given clustering is the average predictive error of the global. This measure also serves to select the often unknown parameter of the number of clusters.

In this talk, we will introduce the idea of using global models for clustering time series, showcasing several algorithms and results on simulation and real datasets. The time series models include the classical linear autoregressive family, but also neural networks and decision trees. Finally, we will draw connections between global models and classic algorithms such as k-means and discuss implicit limitations of the approach.

Keywords: clustering, time series, global models, data pooling, autoregression

Pablo Montero-Manso
University of Syndey, Australia e-mail: pablo.monteromanso@syndey.edu.au

Ángel López-Oriona
University of A Coruña, Spain

José A. Vilar
University of A Coruña, Spain

Clustering and Classifying Time Series in the Sktime Toolkit: a Practical Review of Latest Advances in the Field

Anthony Bagnall

sktime (https://github.com/alan-turing-institute/sktim)is an open source, scikit learn compatible, toolkit for machine learning with time series. It was conceived in 2019 via a collaboration with the Alan Turing Institute and has matured through the growth of a vibrant open source community. It contains modules for a range of learning tasks such as forecasting, annotation and transformation. Researchers at UEA have played a key role in researching and implementing algorithms for time series classification (TSC) and clustering (TSCL) within the sktime framework. I will present a practical overview of the TSC and TSCL functionality available within sktime, with specific emphasis on our new classification algorithm, HIVE-COTEv2.0 [1] (HC2). HC2 is a meta ensemble of four classifiers, built on four different data representations. It is state of the art for both univariate [2] and multivariate TSC [3]. I will also demonstrate how to develop a simple pipeline classifier and compare performance to our published results. The clustering module is a more recent addition to sktime. I will present some recent experimental clustering benchmark results and show how sktime can be used with other toolkits such as tslearn to perform TSCL.

Keywords: time series classification, time series clustering, sktime

References

1. Middlehurst M., Large J., Flynn M., Lines J., Bostrom A. and Bagnall A: HIVE-COTE 2.0: a new meta ensemble for time series classification. Machine Learning **110**, 3211-3243 (2021)
2. Bagnall A, Lines J., Bostrom A., Large J. and Keogh E. Middlehurst M., Large J., Flynn M., and : Data Mining and Knowledge Discovery **31**, 606–660 (2017)
3. Ruiz A.P., Flynn M, Large J., Middlehurst M., Bagnall A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **35(2)**, 401–449 (2021)

Anthony Bagnall
University of East Anglia, Norwich, UK e-mail: ajb@uea.ac.uk

Effect of Type 2 Diabetes and its Genetic Susceptibility on Severity and Mortality of COVID-19 in UK Biobank

Aeyeon Lee, Youngkwang Cho, Jun Li, Taesung Park, Wonil Chung, and Liming Liang

Although Type 2 diabetes (T2D) have been known as one of the important risk factors for the severity and mortality of COVID-19, the effect of T2D and its genetic susceptibility on COVID-19 are largely unknown. We analyzed the population-based cohort data of 459, 188 individuals from UK Biobank with COVID-19 test results, individuals' hospitalization data and death-related records during the period from March 11, 2020 to December 20, 2021. First, we investigated the association of T2D, and its genetic susceptibility with COVID-19 infection using multivariable logistic regression model. To capture overall genetic susceptibility for T2D, we computed polygenic risk scores (PRS) based on summary statistics from UK Biobank. In the multivariable logistic models, we found that the odds ratio (OR) of T2D was 1.555 ($P = 3.49\times10^{-86}$) and OR of PRS for T2D with one-unit (= standard deviation) increase in PRS was 1.064 ($P = 3.11 \times 10^{-12}$), indicating the roles of T2D-related genetics in the pathogenesis of COVID-19 infection. Next, we performed multivariable Cox proportional hazard models to investigate the effect of T2D patients infected with COVID-19 on the survival times. The estimated survival curves and pairwise log-rank tests showed that the estimated hazard for COVID-19 infected T2D patients were 4.67 times ($P = 9.88\times10^{-324}$) and 2.58 times ($P = 6.20\times10^{-231}$) higher than individuals without COVID-19 infection and T2D, respectively and the hazard ratio (HR) of PRS for T2D with one-unit increase in PRS was 1.088 ($P = 4.76 \times 10^{-14}$). Furthermore, we found the mortality of COVID-19 infected T2D patients was dramatically increased compared to T2D patients not infected with COVID-19 and the mortality of individuals with high genetic susceptibility for T2D was increased as well.

Aeyeon Lee
Department of Statistics and Actuarial Science, Soongsil University, Seoul, 06978, Korea

Youngkwang Cho
Department of Statistics and Actuarial Science, Soongsil University, Seoul, 06978, Korea

Jun Li
Department of Nutrition, Harvard TH Chan School of Public Health, Boston, MA 02115, USA

Taesung Park
Department of Statistics, Seoul National University, Seoul, 08826, Korea

Wonil Chung
Department of Statistics and Actuarial Science, Soongsil University, Seoul, 06978, Korea

Liming Liang
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Prognosis of COVID-19 Patients by the Underlying Diseases and Drug Treatment in Korea

Ho Kim and Taerim Lee

Abstract: Certain underlying diseases such as diabetic mellitus and hypertension are a risk factor for the severity and mortality of coronavirus disease (COVID-19) patients. Furthermore, both angiotensin converting enzyme inhibitors (ACEi) and angiotensin II receptor blockers (ARBs) are controversial at role in the process of COVID-19 cases. The aim of the study was to investigate whether underlying diseases and taking ACEi/ARBs, affect the duration of hospitalization and mortality in patients with confirmed COVID-19. Among the comorbidities, a history of hypertension (hazard ratio [HR], 1.51; 95% confidence interval [CI], 1.056–2.158) and diabetes (HR, 1.867; 95% CI, 1.408–2.475) were associated significantly with mortality. Furthermore, heart failure (HR, 1.391; 95% CI, 1.027–1.884), chronic obstructive pulmonary disease (HR, 1.615; 95% CI, 1.185–2.202), chronic kidney disease (HR, 1.451; 95% CI, 1.018–2.069), mental disorder (HR, 1.61; 95% CI, 1.106–2.343), end stage renal disease (HR, 5.353; 95% CI, 2.185–13.12) were also associated significantly with mortality. The underlying disease has increased the risk of mortality in patients with COVID-19. Diabetes, hypertension, cancer, chronic kidney disease, heart failure, and mental disorders increased mortality. Controversial whether taking ACEi/ARBs would benefit COVID-19 patients, in our study, patients taking ACEi/ARBs had a higher risk of mortality.

Keywords: COVID-19; underlying disease; medical treatment

References

1. CDC COVID-19 Response Team; Chow, N.; Fleming-Dutra, K.; Gierke, R.; Hall, A.; Hughes, M.; Pilishvili, T.; Ritchey, M.; Ritchey, K.; Skoff, T.; et al. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—United States, February 12–March 28, 2020. Morb. Mortal. Wkly. Rep. 2020, 69, 382.

Ho Kim
Institute of Health and Environment and Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul 08826, Korea, e-mail: platin@snu.ac.kr

Taerim Lee
Department of Statistics & Data Science, Korea National Open University, Seoul 03087, Korea, e-mail: trlee691@gmail.com

Algorithms for Clustering COVID-19 Data: a Holistic Overview of Current Trends and New Visual Approaches

Eun-Kyung Lee

COVID 19 was first discovered in China, and it spread worldwide and became a pandemic. Since then, many researchers have tried to analyze the temporal tracking of cases and death of COVID-19. In this talk, we focused on the predictive models and clustering methods of the time series of growth rates regarding confirmed cases and deaths of COVID-19. We reviewed the clustering method of time series data included in the papers since 2020. We compared various methods extensively, including the classic statistical methods (k-means, PCA, factor analysis, etc.), fuzzy time series models, functional data analysis models, and deep learning models. We also overviewed the visual approaches to the time course data and proposed new visual approaches. We applied these clustering and visualization methods to the cases and death of COVID-19 in each country for comparison.

Keywords: COVID-19, time series, cluster analysis, visualization

Eun-Kyung Lee
Ewha Womans University, Department of Statistics, Seoul, Republic of Korea
e-mail: lee.eunk@ewha.ac.kr

Embedded Word MCA Biplots for Sentiment Visualisation: Application to COVID-19 Related Tweets

Zoë-Mae Adams, Johané Nienkemper-Swanepoel, Niël le Roux, and Sugnet Lubbe

Social media platforms are continually gaining popularity which results in vast amounts of shared data in the form of images, videos and text. Twitter is a micro-blogging platform which allows the sharing of short messages that are labelled according to a specific key word (i.e. *tag*), representing a relevant topic or theme. These messages reflect personal opinions with subjective content which could provide insight to grasp the underlying attitude towards specific topics. During the global COVID-19 pandemic users could easily share messages by using social media platforms containing information on for example regulations on lockdown or vaccination. Twitter's application programming interface (API) allows the procurement of posts made on the platform for a specific Twitter tag, timeframe and location within a specified radius. In this study the unstructured pieces of text, Tweets, are processed and the sentiment of the remaining words are classified using two lexicons. Multi-dimensional visualisation enables the exploration of the associations between the Twitter users based on the resultant sentiment scores of their posts. A multiple correspondence analysis (MCA) biplot is embedded with the extracted words to enable the simultaneous interpretation of the underlying sentiment of the processed Tweets. This paper presents two case studies of COVID-19 related Tweets. The first case study considers posts made by South African users in three cities (Cape Town, Johannesburg and Durban), with the second case study evaluating the sentiment towards COVID-19 on a global scale by considering three predominantly English-speaking countries (South Africa, Australia and United Kingdom).

Keywords: biplots, covid-19 tweets, multiple correspondence analysis, sentiment classification, web scraping

Zoë-Mae Adams
Centre for Multi-Dimensional Data Visualisation (MuViSU), Department of Statistics and Actuarial Science, Stellenbosch University, South Africa, e-mail: MUVISU@sun.ac.za

A General Framework for Implementing Distance Measures for Categorical Variables

Carlo Cavicchia, Michel van de Velden, Alfonso Iodice D'Enza, and Angelos Markos

In many statistical methods, distance plays an important role. For instance, data visualization, classification and clustering methods require quantification of distances among objects. How to define such distance depends on the nature of the data and/or problem at hand. For distance between numerical variables, in particular in multivariate contexts, there exist many definitions that depend on the actual observed differences between values. It is worth underlining that often it is necessary to rescale the variables before computing the distances. Many distance functions exist for numerical variables, see [2] for a detailed list. For categorical data, defining a distance is even more complex as the nature of such data prohibits straightforward arithmetic operations. Specific measures therefore need to be introduced that can be used to describe or study structure and/or relationships in the categorical data. In this paper, we introduce a general framework that allows an efficient and transparent implementation for distance between categorical variables. We show that several existing distances (for example a distance measure proposed by Ahmad and Dey [1] that incorporates association among variables) can be incorporated into the framework. Moreover, our framework quite naturally leads to the introduction of new distance formulations as well.

Keywords: categorical data, distance, cluster analysis

References

1. Ahmad, A. and Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, **63**, 2, 503–3527 (2007)
2. Mardia, K.V.: Some properties of classical multidimesional scaling. Communications in Statistics - Theory and Methods, **7**, 13, 1233–1241 (1978)

C. Cavicchia · M. van de Velden
Econometric Institute, Erasmus University Rotterdam,
e-mail: {cavicchia,vandevelden}@ese.eur.nl

A. Iodice D'Enza
Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II
e-mail: iodicede@unina.it

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

Some Descriptive Statistics of Aggregated Symbolic Data

Junji Nakano, Nobuo Shimizu, and Yoshikazu Yamamoto

When we have a huge amount of data, we sometimes are interested in comparing meaningful groups of data, not individual observations. Aggregated symbolic data (ASD) expresses a group of observations that have continuous and categorical variables by using up to second moments of variables. ASD for a group of data is equivalent to the set of means, variances, and correlations for continuous variables, Burt matrix for categorical variables, and means of a continuous variable against one value of a categorical variable. As ASD for many categorical variables is still complicated, it is preferable to have simple measures of location and dispersion for a categorical variable, and a measure of the correlation between two categorical and/or continuous variables. We propose such measures by specifying appropriate scores to categorical values. They are compared with measures that are defined as extensions of the polychoric correlation coefficient [1].

Keywords: categorical variable, continuous variable, measure of correlation, measure of dispersion, measure of location

References

1. Olsson, U.: Maximum Likelihood Estimation of the Polychoric Correlation Coefficient, Psychometrika, **12**, 443–460 (1979)

Junji Nakano
Chuo University, Tokyo, Japan. e-mail: nakanoj@tamacc.chuo-u.ac.jp

Nobuo Shimizu
The Institute of Statistical Mathematics, Tokyo, Japan, e-mail: nobuo@ism.ac.jp

Yoshikazu Yamamoto
Tokushima Bunri University, Kagawa, Japan. e-mail: yamamoto@wjasp.bunri-u.ac.jp

Variable Screening in High Dimensional Regression via Random Projection Ensembles

Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, and Angela Montanari

Random projections (RP) are a fairly new tool for dimension reduction employed in several multivariate data analysis contexts (see, e.g. [1, 2]). In this paper, we present a novel approach based on RP ensemble for variable screening in the multiple linear regression framework. By employing axis-aligned random projections, column sub-sampling is performed, thus constituting an even cheaper way of randomized dimension reduction outside the class of Johnson-Lindenstrauss transforms. Differently from the approach proposed in [3], the method allows to account for the correlation among the predictors, and returns a variable ranking based on their importance.

We provide numerical results based on synthetic and real data as well as basic theoretical results that characterize the proposed solution.

Keywords: variable screening, random projections, high-dimensional linear regression

References

1. Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. B **70**, 849–911 (2008)
2. Gataric, M., Wang, T. and Samworth, R. J.: Sparse principal component analysis via axis-aligned random projections. J. R. Stat. Soc. B **82**, 329–359 (2020)
3. Thanei, G.A., Heinze, C., Meinshausen, N.: Random Projections for Large-Scale Regression. In: Ahmed, S. (eds) Big and Complex Data Analysis, pp. 51-68. Contributions to Statistics. Springer, Cham (2017).

Laura Anderlucci
Department of Statistical Sciences - University of Bologna, Italy,
e-mail: laura.anderlucci@unibo.it

Matteo Farnè
Department of Statistical Sciences - University of Bologna, Italy,
e-mail: matteo.farne@unibo.it

Giuliano Galimberti
Department of Statistical Sciences - University of Bologna, Italy,
e-mail: giuliano.galimberti@unibo.it

Angela Montanari
Department of Statistical Sciences - University of Bologna, Italy,
e-mail: angela.montanari@unibo.it

Model-based Clustering and Dimension Reduction for Multidimensional Social Networks

Michael Fop, Silvia D'Angelo, and Marco Alfò

Social network data are relational data recorded among a group of actors, interacting in different contexts. Often, the same set of actors can be characterized by multiple social relations, captured by a multidimensional network. A common situation is that of colleagues working in the same institution, whose social interactions can be defined on professional and personal levels. In addition, individuals in a network tend to interact more frequently with similar others, naturally creating communities. Latent space models for network data are useful to recover clustering of the actors, as they allow to represent similarities between them by their positions and relative distances in an interpretable low dimensional social space. We propose the infinite latent position cluster model for multidimensional network data, which enables dimension reduction and model-based clustering of actors interacting across multiple social dimensions. The model is formulated within a Bayesian nonparametric framework, which allows to perform automatic inference on the clustering allocations, the number of clusters, and the latent social space.

Keywords: Bayesian nonparametric, latent space model, model-based clustering, statistical network analysis

Michael Fop
School of Mathematics and Statistics, University College Dublin Ireland, Ireland
e-mail: michael.fop@ucd.ie

Silvia D'Angelo
School of Mathematics and Statistics, University College Dublin Ireland, Ireland

Marco Alfò
Department of Statistics, Sapienza University of Rome, Italy

Conditional Gaussian Mixture Modeling

Volodymyr Melnykov and Yang Wang

Due to a potentially high number of parameters, finite mixture models are often at the risk of overparameterization even for a relatively low number of components. This can lead to overfitting and result in mixture order underestimation. One of the most popular approaches to alleviate this issue is to reduce the number of parameters by considering parsimonious models. The vast majority of techniques in this area focus on the reparameterization of covariance matrices associated with mixture components. We propose an alternative approach that shows great modeling flexibility. The utility of the proposed methodology is demonstrated on simulated as well as well-known classification data sets.

Keywords: finite mixture model, model-based clustering, parsimonious models

Volodymyr Melnykov
The University of Alabama, Tuscaloosa, AL 35487, USA, e-mail: vmelnykov@ua.edu

Yang Wang
College of Charleston, Charleston, SC 29424, USA, e-mail: wangy4@cofc.edu

Classification Over Text, Relational Databases and Graphs - Software and Case Studies

Tomáš Kliegr

Best performing methods often produce models that are hard to interpret, leading to the so-called accuracy-interpretability trade-off. Also, each data type typically requires a different type of model, such as BERT transformers for text or node embeddings for graphs. In projects involving multiple modalities, this leads to a mix of opaque models, an interpretability and interoperability Babylon.

This talk will cover rule-based methods as a possible "white-box" Swiss knife applicable to multiple data types, including tabular data, text and even large knowledge graphs with millions of edges and nodes.

Given the advances in model-agnostic explanation algorithms, do rule models still have an edge in interpretability over the more opaque classification workhorses such as random forests? The talk will hint at answers through use cases worked on at DIKE, such as comparing rule-based explanations with LIME and Shapley plots in the context text-mining of research articles on COVID-19 [1]. We will also cover tools such as the *arc* R package for rule-based classification of tabular data [2], the Action rule mining system [3], the cloud-based *EasyMiner* rule classifier and editor [4], and the *RDFRules* rule learner for knowledge graphs [5].

Keywords: knowledge graphs, rules, explainable machine learning

Acknowledgements Presented research was partly supported by support of the CIMPLE project (CHIST-ERA-19-XAI-003) and VSE IGA 40/2021.

References

1. Beranová, L., Joachimiak, M. P., Kliegr, T., Rabby, G., and Sklenák, V. Why was this cited? explainable machine learning applied to covid-19 research literature. *Scientometrics* (2022), 1–37.
2. Hahsler, M., Johnson, I., Kliegr, T., and Kuchař, J. Associative classification in r: arc, arulesCBA, and rCBA. *R Journal 9*, 2 (2019).
3. Sýkora L., and Kliegr, T. Action Rules: Counterfactual Explanations in Python. *RuleML Challenge* (2020).
4. Vojíř, S., Zeman, V., Kuchař, J., and Kliegr, T. Easyminer.eu: Web framework for interpretable machine learning based on rules and frequent itemsets. *Knowledge-Based Systems 150* (2018), 111–115.
5. Zeman, V., Kliegr, T., and Svátek, V. RDFrules: Making RDF rule mining easier and even more efficient. *Semantic Web 12*, 4 (2021), 569–602.

Tomáš Kliegr
Department of Information and Knowledge Engineering (DIKE), Faculty of Informatics and Statistics, Prague University of Economics and Business, W Churchill sq. 4, Prague, Czech Republic, e-mail: tomas.kliegr@vse.cz.

Towards Deep and Interpretable Rule Learning

Johannes Fürnkranz

Inductive rule learning is concerned with the learning of classification rules from data. Learned rules are inherently interpretable and easy to implement, so they are very suitable for formulating learned models in many domains. Nevertheless, current rule learning algorithms have several shortcomings. First, with respect to the current praxis of equating high interpretability with low complexity, we argue that while shorter rules are important for discrimination, longer rules are often more interpretable than shorter rules, and that the tendency of current rule learning algorithms to strive for short and concise rules should be replaced with alternative methods that allow for longer concept descriptions. In general, the mere syntactic comprehensibility of the learned concepts does often not yield convincing or plausible rules, and factors such as semantic coherence or the a priori relevance of used conditions should be explicitly encoded as objectives in an interpretable rule learning algorithm. Human cognitive biases can be one possible road towards the design of an interpretability bias for rule learning [3]. Second, we think that the main impediment of current rule learning algorithms is that they are not able to learn deeply structured rule sets, unlike the successful deep learning techniques. Both points are currently under investigation in our group, and we will show some preliminary results [1, 2].

Keywords: inductive rule learning, interpretability, explainable AI, deep learning

References

1. Beck, F., Fürnkranz, J.: An empirical investigation into deep and shallow rule learning. Frontiers Artif. Intell. **4**:689398 (2021)
2. Beck, F., Fürnkranz, J., Quoc, P.H.V.: Structuring rule sets using binary decision diagrams. In: Proceedings of the 5th International Joint Conference on Rules and Reasoning (RuleML+RR), Leuven, Belgium, pp. 48–61. Springer (2021)
3. Fürnkranz, J., Kliegr, T., Paulheim, P.: On cognitive preferences and the plausibility of rule-based models. Mach. Learn. **109**(4):853–898 (2020)

Johannes Fürnkranz
Institute for Application-Oriented Knowledge Processing (FAW), Johannes-Kepler Universität, Linz, Austria, e-mail: juffi@faw.jku.at

Current Challenges in Interpretable Machine Learning and Partitioning Approaches

Bernd Bischl

Model-agnostic interpretation methods in machine learning produce explanations based on non-linear, non-parametric prediction models. Explanations are often represented in the form of summary statistics or visualizations, e.g., feature importance values or effects. Many interpretation methods either describe the behavior of a black-box model locally for a specific observation or globally for the entire model and input space. Methods that produce regional explanations and lie between local and global explanations are rare and not well studied, but offer a flexible way to combine advantages of both types of explanations. Here, we will focus on subgroup approaches for IML methods, where interpretable areas in the input space are often induced by a combination of recursive partitioning and IML. These subgroup approaches will be studied in the contexts of interpretable permutation feature importance and PDPs [1], interaction detection [2], and interpretable hyperparameter tuning [3].

Keywords: machine learning, interpretable machine learning, xai

References

1. Molnar, C., König, G., Bischl, B., & Casalicchio, G. (2020). Model-agnostic Feature Importance and Effects with Dependent Features–A Conditional Subgroup Approach. arXiv preprint arXiv:2006.04628.
2. Herbinger, J., Bischl, B., & Casalicchio, G. (2022, May). REPID: Regional Effect Plots with implicit Interaction Detection. In International Conference on Artificial Intelligence and Statistics (pp. 10209-10233). PMLR.
3. Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., & Bischl, B. (2021). Explaining Hyperparameter Optimization via Partial Dependence Plots. Advances in Neural Information Processing Systems, 34.

Bernd Bischl
Department of Statistics, LMU, Ludwigstraße 33, D-80539 München,
e-mail: bernd.bischl[at]stat.uni-muenchen.de

Part VI
Benchmarking Challenge

Vine Copula Mixture Models and Clustering for Non-Gaussian Data

Özge Sahin and Claudia Czado

The majority of finite mixture models suffer from not allowing asymmetric tail dependencies within components and not capturing non-elliptical clusters in clustering applications. Since vine copulas are very flexible in capturing these dependencies, a novel vine copula mixture model for continuous data is proposed. The model selection and parameter estimation problems are discussed, and further, a new model-based clustering algorithm is formulated. The use of vine copulas in clustering allows for a range of shapes and dependency structures for the clusters. The simulation experiments illustrate a significant gain in clustering accuracy when notably asymmetric tail dependencies or/and non-Gaussian margins within the components exist. The analysis of real data sets accompanies the proposed method. The model-based clustering algorithm with vine copula mixture models outperforms others, especially for the non-Gaussian multivariate data.

Keywords: dependence, ECM algorithm, model-based clustering, multivariate finite mixtures, pair-copula

Özge Sahin
Department of Mathematics, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany, e-mail: ozge.sahin@tum.de

Claudia Czado
Department of Mathematics, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany, e-mail: cczado@ma.tum.de

Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering

Zdeněk Šulc and Hana Řezanková

This contribution examines 13 similarity measures for data characterized by nominal variables in hierarchical clustering. Most of the measures come from [1], where they were initially studied in outlier detection tasks, and two of them are newly proposed in [2]. The inspected measures consider additional characteristics of the clustered dataset, such as a frequency distribution of categories or the number of categories of a given variable, which should lead to a better cluster quality than the commonly used simple matching approach. The experiment is conducted on 60 generated datasets. It compares and evaluates the similarity measures regarding the quality of the produced clusters in hierarchical clustering using the mean ranked scores of two internal evaluation criteria. The calculations are performed using the `nomclust` R package [3]. The obtained results determine which similarity measures are the most suitable for use with a given number of variables or a linkage algorithm.

Keywords: similarity measures, nominal variables, hierarchical clustering

References

1. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the eighth SIAM International Conference on Data Mining, pp. 243–254. (2008)
2. Šulc, Z., Řezanková, H. Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering. J Classif **36**, 58–72. (2019)
3. Šulc, Z., Cibulková, J., Řezanková, H. Nomclust 2.0: an R package for hierarchical clustering of objects characterized by nominal variables. Comput Stat (2022)

Zdeněk Šulc
Department of Statistics and Probability, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 130 67 Prague 3, Czechia, e-mail: `zdenek.sulc@vse.cz`

Hana Řezanková
Department of Statistics and Probability, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 130 67 Prague 3, Czechia, e-mail: `hana.rezankova@vse.cz`

Comparing Model Selection Techniques to Determine the Number of Overlapping Clusters for the Additive Profile Clustering Model

Tom F. Wilderjans, Julian Rossbroich, and Jeffrey Durieux

In several areas of science, researchers aim at disclosing the mechanisms that generated object by variable data, like, for example, patient by symptom or consumer by brand data. Regularly, based on previous knowledge, it makes sense to assume that these mechanisms can be nicely captured through an object clustering. In such a case, researchers very often opt for a partitioning method, like, for example, k-means, which results in non-overlapping clusters. Sometimes, however, expectations are that objects can be grouped into clusters that overlap, implying that an object may belong to multiple clusters. Applied to the patient by symptom data, for example, in which clusters correspond with syndromes, it is quite natural to assume that patients may suffer from multiple syndromes at the same time as co-morbidity of syndromes is often observed in clinical practice. To extract the overlapping object clusters, Mirkin's additive profile (overlapping) clustering model may be used. A non-trivial task consists of determining the optimal number of overlapping clusters that are present in a given empirical data set. Up to now, however, although some methods for model selection for ADPROCLUS were proposed before, no systematic attempt of investigating this issue of model selection has been undertaken. Therefore, in this presentation, we evaluate in an extensive simulation study several model selection techniques. In partiular, several new model selection methods for ADPROCLUS, with some of them being methods for the partitioning case tailored to the context of overlapping clustering (e.g., AIC, CH-index) are compared to existing methods (e.g., CHull, cross-validation). As such, in order to build a cumulative body of knowledge, our study is a benchmarking study in which the performance of new methods is carefully compared to the performance of existing methods.

Keywords: overlapping clustering, additive profile clustering, model selection, simulation study, benchmarking

Tom F. Wilderjans · Jeffrey Durieux
Faculty of Social and Behavioral Sciences, Leiden University, Leiden, the Netherlands
e-mail: {t.f.wilderjans,j.durieux}@fsw.leidenuniv.nl

Julian Rossbroich
Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland
e-mail: julian.rossbroich@fmi.ch

Pitfalls of Automatic Optimization Procedures and Benchmarking in Cluster Analysis

Quirin Stier and Michael C. Thrun

The pitfalls and challenges of automatic approaches are outlined in the case that relevant and possibly prior unknown relationships in high-dimensional biological datasets are to be discovered [1]. Priorly, [2] proposed one or more unsupervised quality measures for the automatic selection of clustering algorithms and their parameter optimization. However, employing optimization procedures within automated pipelines is biased and not recommended if we assume there may be only one optimal partitioning of data, e.g., diagnoses or therapies [1]. Thus, the limitations of a clustering algorithm induced by a global clustering criterion cannot be overcome by optimizing the algorithm parameters which only reduces the variance but not the intrinsic bias of the criterion [1]. Furthermore, such optimization can yield significant improvements even if the dataset does not possess any cluster structure. Finally, our work shows that benchmarking clustering algorithms using first-order statistics or box plots on a small number of trials leads to misleading comparisons between algorithms. Assuming patterns in the data which can be recognizd by experts, we use artifical generated datasets [3]. On these datasets, 41 open-source and state-of-the-art algorithms standardized within R and Python in the "FCPS" library [4] are evaluated to disprove the claim of [2] that automatic algorithm and parameter selection by unsupervised quality measures is a viable approach in cluster analysis.

Keywords: cluster analysis, benchmarking, quality measure

References

1. Thrun, M. C. Distance-based clustering challenges for unbiased benchmarking studies. Nature Scientific Reports 11, 18988, doi:10.1038/s41598-021-98126-1 (2021)
2. Wiwie, C., Baumbach, J., Röttger, R. Comparing the performance of biomedical clustering methods. Nature Methods 12, 1033 (2015)
3. Thrun, M. C., Ultsch, A. Clustering benchmark datasets exploiting the fundamental clustering problems. Data in Brief 30, 105501, doi:10.1016/j.dib.2020.105501 (2020)
4. Thrun, M. C., Stier, Q. Fundamental clustering algorithms Suite SoftwareX 13, 100642, doi:10.1016/j.softx.2020.100642 (2021)

Quirin Stier
IAP-GmbH Intelligent Analytics Projects, Adelheidsdorf, Germany,
e-mail: q.stier@iap-gmbh.de

Michael C. Thrun
Dept. of Mathematics and Computer Science, Philipps-University, Marburg, Germany,
e-mail: m.thrun@informatik.uni-marburg.de

Part VII
Contributed Abstracts

Biplots for Categorical and Ordinal Data Based on Logistic Responses

Jose Luis Vicente-Villardon

A joint representation of individuals and variables in a data matrix is called a Biplot. Biplots were proposed 50 years ago in [1].

When variables are binary, nominal or ordinal, a classical linear biplot representation is not adequate. More recently, biplots for categorical data, based logistic response models, have been proposed for binary [2], or nominal data [3]. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The methods are related to logistic regression in the same way as Classical Biplots are related to linear regression, thus are referred as Logistic Biplots. In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. The geometry of those kinds of biplots for binary, nominal or ordinal data is studied.

For binary data we obtain straight lines as representations of the variables.

For nominal data the representation of the variables on the biplot is not a straight line but a "prediction region" and for ordinal data a straight line is obtained if the "proportional odds" model is used.

Algorithms for the construction on the biplots based on gradient descent methods are also provided.

The applicability and interpretation of the logistic biplots is illustrated with several real data applications.

Keywords: biplot, categorical data, logistic biplot

References

1. Gabriel, K. R. : The biplot graphic display of matrices with application to principal component analysis. Biometrika, **58** (3), 453-467. (1971).
2. Vicente-Villardon, J.L., Galindo, M.P., Blazquez-Zaballos, A.: Logistic biplots. In: Greenacre, M., Blasius, J. (eds.) Multiple Correspondence Analysis and related methods, pp. 503-521. Chapman and Hall, New York (2006)
3. Hernandez-Sanchez, J. C., Vicente-Villardon, J. L. (2017). Logistic biplot for nominal data. Advances in Data Analysis and Classification, **11** (2) 307-326.

Jose Luis Vicente-Villardon
Departamento de Estadistica, Universidad de Salamanca,Spain e-mail: villardon@usal.es

The Biplot Inner Product for Interpretation and Derivation of Eigenvector Methods

Cajo J. F. ter Braak

To celebrate K. Ruben Gabriel's biplot paper [1], I describe its influence, via my supervisor Leo C. A. Corsten, on my first to last papers (1981-2021) and on my teaching. Being not just a plot of two sets of items, Gabriel's biplot shifted attention from the meaning of individual PCA axes to inference from the first 2-3 axes together via the inner product interpreted geometrically, namely as the product of the arrow lengths and the cosine of their angle or as the product of the lengths of the one arrow and the other projected on to it. These simple equations cannot only be used for interpreting biplots, I will show that they can also be used to derive the eigen equations. Whereas later, even non-linear, extensions turned the arrows in to calibrated, possibly curved, lines, I always promoted inference using the rank order of the projection points.
Gabriel showed that the biplot could be used beyond PCA, for example, in canonical correlation analysis and variants thereof [2]. Double-constrained correspondence analysis even allows for plots with four sets of items, pairs of which approximated different summary statistics of the method [3].
I will also briefly describe how the distinction between predictive and interpolative biplot was discovered and how multivariate analysis was taught in the 70-80s in France and Japan, compared to the English literature that was focussing more on weighted least-squares approximation and Gabriel's biplot interpretation of factorial diagrams.

Keywords: biplot, inner product, pca, canonical correlation analysis, double constrained correspondence analysis

References

1. Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**, 453–467 (1971)
2. ter Braak, C.J.F.: Interpreting canonical correlation analysis through biplots of structural correlations and weights. Psychometrika. **55**, 519–531 (1990)
3. ter Braak, C.J.F., Smilauer, P., Dray, S.: Algorithms and biplots for double constrained correspondence analysis. Environ. Ecol. Stat. **25**, 171–197 (2018)

Cajo J. F. ter Braak
Biometris, Wageningen University & Research, Wageningen, the Netherlands
e-mail: cajo.terbraak@wur.nl

Fifty Years of Biplots: Some Remaining Enigmas and Challenges

Jan Graffelman

Biplots have found applications in many fields of science, where they are often used to detect groups, outliers or other regularities in multivariate data. A recent search at the web of science (using the core collection) reveals there are well over 2,200 scientific articles that refer to biplots since the term was first coined by Gabriel [1]. Several textbooks on biplots are available with a varying level of mathematical depth, though classical textbooks on multivariate analysis have been slow at incorporating the concept, or at recognising its universal value for all classical methods that are based on the singular value decomposition.

In applied scientific studies, several aspects of biplot construction and interpretation are still not well understood, to the point that it is easy to find a poor biplot in a high-impact scientific journal. Some of the more problematic aspects will be addressed in the talk and concern: the choice of the multivariate method, aspect ratio, biplot scaling, interpretation rules and (sub)optimality of approximations among others. There is a lot of software available for making biplots, but in spite of this, the truth is that many users would not be able to make an optimal biplot for representing one of the most elementary matrices: the correlation matrix.

The growing size of datasets being analysed poses an important challenge to biplot methodology. For example, by 2015 the 1,000 genomes project contained information on 88 million genetic polymorphisms for over 2,500 human individuals, obviously impossible to sensibly represent in a single biplot. In this situation, cases and variables are finally often presented in separate plots, thereby sacrificing the biplot's most appealing feature: its ability to *jointly* represent and interpret observations and variables. We address some examples in the high-dimensional context, where biplots are too dense, often have low goodness-of-fit, and where aggregation, clustering or filtering are needed to make their application feasible.

Keywords: biplot, singular value decomposition, high-dimensional data

References

1. Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika. **58**(3), 453–467 (1971)

Jan Graffelman
Technical University of Catalonia, Carrer Jordi Girona, 1-3, 08034, Barcelona, Spain
e-mail: jan.graffelman@upc.edu

Outlier Detection for BIG Functional Data

Rosa E. Lillo, Oluwasegun T. Ojo, and Antonio Fernández-Anta

The need to find influential users in social networks motivates a multidisciplinary line of research whose final result, from the point of view of Statistics, is the development and implementation of several procedures for detecting outliers in functional data that are scalable for massive data and that are also competitive, in terms of performance, with the most used algorithms in the usual literature on functional data. Brushstrokes of theoretical contributions and various fields of practical application will be provided.

Keywords: functional data, outlier detection, scalable algorithms

References

1. Azcorra, A, Cuevas R., Chiroque, L., Fernández A., Laniado, H. Lillo, R.E, Romo J, Sguera, C.: Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks. Scientific Reports **8**, 6995 Nature, https://www.nature.com/articles/s41598-018-24874-2 (2018)
2. Dai, W. and Genton, M. G.: Multivariate functional data visualization and outlier detection. Journal of Computational and Graphical Statistics, **27(4)**:923–934. (2018)
3. Huang, H. and Sun, Y. A decomposition of total variation depth for understanding functional outliers. Technometrics, **61(4)** 445–458.(2019)
4. Ojo, O. Fernández-Anta, A, Lillo, R. E., Sguera, C: Detecting and classifying outliers in big functional data. Advances in Data Analysis and Classification https://doi.org/10.1007/s11634-021-00460-9, (2021)

Rosa E. Lillo
uc3m-Santander Big Data Institute, Department of Statistics, Universidad Carlos III de Madrid, Spain, e-mail: rosaelvira.lillo@uc3m.es

Oluwasegun T. Ojo
IMDEA Networks, Spain, e-mail: oluwasegun.ojo@imdea.org

Antonio Fernández-Anta
IMDEA Networks, Spain, e-mail: antonio.fernandez@imdea.org

Outlier and Novelty Detection for Functional Data: a Semiparametric Bayesian Approach

Francesco Denti, Andrea Cappozzo, and Francesca Greselin

A novelty detection model can be seen as a supervised classifier, trained on a fully-labeled training set, that allows for the presence of new classes in the test set not previously observed among the training units. When dealing with functional data, this requires learning the main patterns for the curves in the known classes, whilst being able to isolate signals that possess distinctive characteristics in the unlabeled set. In order to tackle this challenging problem, we propose a two-stage Bayesian semi-parametric novelty detector [2]. In the first stage, robust estimates are extracted from the training set via the Minimum Regularized Covariance Determinant (MRCD) estimator [1]. In the second stage, such information is employed to elicit informative priors within a Bayesian mixture of known groups plus a novelty term. To reflect the lack of knowledge on the latter component, we resort to a Dirichlet Process mixture model, thus overcoming the problematic a-priori specification of the expected number of novelties that may be present in the test set. The described methodology is applied to a spectroscopic dataset within a food authenticity study.

Keywords: bayesian mixture model, dirichlet process mixture model, functional data, minimum regularized covariance determinant

References

1. Denti, F., Cappozzo, A., Greselin, F.,: A two-stage Bayesian semiparametric model for novelty detection with robust prior information. Stat. Comput. **31**, 42 (2021)
2. Boudt, K., Rousseeuw, P.J., Vanduffel, S.,Verdonck, T.: The minimum regularized covariance determinant estimator. Stat. Comput. **30**, 113–128 (2020)

Francesco Denti
Università Cattolica del Sacro Cuore, Largo Agostino Gemelli, 1, 20123 Milano
e-mail: francesco.denti@unicatt.it

Andrea Cappozzo
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: andrea.cappozzo@polimi.it

Francesca Greselin
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano
e-mail: francesca.greselin@unimib.it

A Geometric Perspective on Functional Outlier Detection

Moritz Herrmann and Fabian Scheipl

Outlier detection in functional data faces specific challenges due to the information-rich and complex nature of functional observations. We [1] consider the problem from a geometric perspective and present a general conceptualization based on the assumption that functional datasets are drawn from a manifold defined by the data's modes of variation in shape, translation, and phase. Theoretical and experimental analyses demonstrate this conceptualization has important advantages. It considerably improves theoretical understanding and allows to describe and analyze complex functional outlier scenarios consistently and in full generality, by differentiating between structurally anomalous outlier data that are off-manifold and distributionally outlying data that are on-manifold, but at its margins. From a practical perspective, we show that well-established manifold learning methods can be used to learn low-dimensional vector-valued representations of functional observations to reliably infer and visualize the geometric structure of functional datasets. Our experiments on synthetic and real data demonstrate that using these representations in combination with the simple outlier scoring method Local Outlier Factors (LOF) yields performances at least on par with existing functional-data-specific methods in a large variety of settings, without the highly specialized, complex methodology and narrow domain of application these methods often entail.

Keywords: functional data analysis, outlier detection, manifold learning, dimension reduction

References

1. Herrmann, M., Scheipl F.: A geometric perspective on functional outlier detection. Stats **4**, 971-1011 (2021)

Moritz Herrmann
Ludwig-Maximilians-University, Department of Statistics, Ludwigstr. 33 80539 Munich
e-mail: moritz.herrmann@stat.uni-muenchen.de

Fabian Scheipl
Ludwig-Maximilians-University, Department of Statistics, Ludwigstr. 33 80539 Munich,
e-mail: fabian.scheipl@stat.uni-muenchen.de

A New Decomposition of Orthogonal Matrices with Application to Common Principal Components

Luca Bagnato and Antonio Punzo

In many statistical problems, the estimation of a $(d \times d)$ orthogonal matrix $\boldsymbol{Q}$ is involved [2]. The orthonormality constraints on $\boldsymbol{Q}$ often makes this estimation difficult. To cope with this problem, we use the well-known PLU decomposition [3], which factorizes any invertible $(d \times d)$ matrix as the product of a $(d \times d)$ permutation matrix $\boldsymbol{P}$, a $(d \times d)$ unit lower triangular matrix $\boldsymbol{L}$, and a $(d \times d)$ upper triangular matrix $\boldsymbol{U}$. Thanks to the QR decomposition [3], we find the formulation of $\boldsymbol{U}$ when the PLU decomposition is applied to $\boldsymbol{Q}$. We call the result as PLR decomposition; it produces a one-to-one correspondence between $\boldsymbol{Q}$ and the $d\,(d-1)\,/2$ entries below the diagonal of $\boldsymbol{L}$, which are advantageously unconstrained real values. Thus, once the decomposition is applied, regardless of the objective function under consideration, we can use any classical unconstrained optimization method to find the minimum (or maximum) of the objective function with respect to $\boldsymbol{L}$. For illustrative purposes, we apply the PLR decomposition in common principle components analysis (CPCA) for the maximum likelihood estimation of the common orthogonal matrix when a multivariate leptokurtic-normal distribution is assumed in each group. Compared to the commonly used normal distribution, the leptokurtic-normal has an additional parameter governing the excess kurtosis [1]; this makes the estimation of $\boldsymbol{Q}$ in CPCA more robust against mild outliers. The usefulness of the PLR decomposition in leptokurtic-normal CPCA is illustrated by two biometric data analyses.

Keywords: orthogonal matrix, matrix decomposition, common principal components, fg algorithm, leptokurtic-normal distribution

References

1. Bagnato, L., and Punzo, A., and Zoia, M.G.: The Multivariate Leptokurtic-Normal Distribution and its Application in Model-Based Clustering. Can. J. Stat. **45**, 95–119 (2017)
2. Graybill, F.A.: An Introduction to Linear Statistical Models. McGraw-Hill (1976)
3. Lütkepohl, H.: Handbook of Matrices. John Wiley & Sons, Chicester (1996)

Luca Bagnato
Università Cattolica del Sacro Cuore, Dipartimento di Scienze Economiche e Sociali, Via Emilia Parmense, 84, 29122 Piacenza, Italia, e-mail: luca.bagnato@unicatt.it

Antonio Punzo
Università di Catania, Dipartimento di Economia e Impresa, Corso Italia, 55, 95129 Catania, Italia, e-mail: antonio.punzo@unict.it

An MML Embedded Approach for Estimating the Number of Clusters

Cláudia Silvestre, Margarida G. M. S. Cardoso, and Mário Figueiredo

Assuming that the data originate from a finite mixture of multinomial distributions, we study the performance of an integrated *Expectation Maximization* (EM) algorithm considering *Minimum Message Length* (MML) criterion to select the number of mixture components. The referred EM-MML approach, rather than selecting one among a set of pre-estimated candidate models (which requires running EM several times), seamlessly integrates estimation and model selection in a single algorithm. Comparisons are provided with EM combined with well-known information criteria – e.g. the Bayesian information Criterion. We resort to synthetic data examples and a real application. The EM-MML computation time is a clear advantage of this method; also, the real data solution it provides is more parsimonious, which reduces the risk of model order overestimation and improves interpretability.

Keywords: finite mixture model, em algorithm, model selection, minimum message length, categorical data

References

1. Figueiredo, M.A.T., Jain, A.K. : Unsupervised Learning of Finite Mixture Models. IEEE T. Pattern Anal. **24**, 381–396 (2002)
2. Novais, L., Faria, S.,: Selection of the number of components for finite mixtures of linear mixed models. J. Int. Math. **24(8)**, 2237–2268 (2021)
3. Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M.: Feature selection for clustering categorical data with an embedded modeling approach. Expert Syst. **32(3)**, 444–453 (2014).

Cláudia Silvestre
Escola Superior de Comunicação Social, Campus de Benfica do IPL 1549-014 Lisboa, Portugal,
e-mail: csilvestre@escs.ipl.pt

Margarida G. M. S. Cardoso
BRU-UNIDE, ISCTE-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal
e-mail: margarida.cardoso@iscte-iul.pt

Mário Figueiredo
Instituto de Telecomunicações, Portugal, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
e-mail: mario.figueiredo@tecnico.ulisboa.pt

Comparison of Segmentation Approaches for Partial Least Squares Path Modeling with Stability Assessment

Sophie Dominique, Mohamed Hanafi, Fabien Llobell, Jean-Marc Ferrandi, and Véronique Cariou

In the social sciences, structural equation modeling has become an established method for analyzing complex interrelationships between manifest and latent variables. In this context, the composite-based Partial Least Squares Path Modeling (PLS-PM) approach [2] has gained popularity over the past decades because of its versatility. Several segmentation methods dedicated to PLS-PM have been proposed to account for the potential heterogeneity of the data [1]. These techniques differ from each other in various aspects such as proceeding in two steps (PLS-PM then segmentation) or not (simultaneous determination of local PLS-PM models per group), being based on finite mixture models, on a distance or more recently on alternate least squares algorithm [1], etc. In this presentation, we propose to compare these segmentation approaches both theoretically according to the criterion they optimize and practically by evaluating the stability of the different segmentations obtained on the basis of a case study pertaining to marketing.

Keywords: partial least squares, clustering, structural equation modeling, marketing

References

1. Fordellone, M., Vichi, M.: Finding groups in structural equation modeling through the partial least squares algorithm. Comput. Stat. and Data Anal. **147**, 106957 (2020)
2. Lohmöller, JB.: Predictive vs. Structural Modeling: PLS vs. ML. In: Latent Variable Path Modeling With Partial Least Squares, pp. 199-226. Springer, Heidelberg (1989)
3. Sarstedt, M.: A review of recent approaches for capturing heterogeneity in partial least squares path modelling. J. of Model. in Manag. **3**, 140-161 (2008)

Sophie Dominique
StatSC, ONIRIS, 44322 Nantes,
Addinsoft, XLSTAT, Paris, France,e-mail: sophie.dominique@oniris-nantes.fr

Mohamed Hanafi · Véronique Cariou
StatSC, ONIRIS, INRAE, 44322 Nantes,
e-mail: mohamed.hanafi@oniris-nantes.fr;veronique.cariou@oniris-nantes.fr

Fabien Llobell
Addinsoft, XLSTAT, Paris, France, e-mail: fllobell@xlstat.com

Jean-Marc Ferrandi
LEMNA, ONIRIS, 44322 Nantes, e-mail: jean-marc.ferrandi@oniris-nantes.fr

Robust Classification for Toroidal Data

Giovanni Saraceno, Luca Greco, and Claudio Agostinelli

Circular data commonly occur in many different fields, such as biology, meteorology and geology, where observations can be measured by angles. Here, we consider the problem of classifying circular observations into one of possible distinct populations and we focus on the situations where observations can be thought as points that lie on the surface of a Torus. Some proposals can be found in literature for the classification of circular data, however this problem is poorly explored in case of toroidal data.

In general, the traditional procedures for the classification problem can be greatly affected by inaccurancies in features and labels of training data. Hence, we propose a procedure based on the weighted likelihood technique which is able to classify new data points scattered on a p-dimensional torus following multivariate Wrapped Normal distributions. In particular, the Weighted CEM algorithm proposed by [1] is applied on the training data set considering the classes separately. Ineed, this estimator is able to handle the model inadequacies in the fitting process by an effective downweighting of observations not following the assumed model. In this way, a pair of robust location and scale estmates are available for each group. In a second step, a set of data-dependent weights is computed for the testing data points for each group-based estimates. Finally, the resulting weights are used to classify each observations into one of the groups or none of them. The finite sample behavior of the proposed procedure is investigated by a Monte Carlo numerical study and real data examples.

Keywords: classification, multivariate wrapped distributions, robust estimators, torus, weighted likelihood

References

1. G. Saraceno and C. Agostinelli and L. Greco (2021). Robust estimation for multivariate wrapped models. *METRON*, **79**:225–240.

G. Saraceno
University of Trento, Trento, Italy, e-mail: giovanni.saraceno@unitn.it

C. Agostinelli
University of Trento, Trento, Italy, e-mail: claudio.agostinelli@unitn.it

L. Greco
University Giustino Fortunato, Benevento, Italy, e-mail: l.greco@unifortunato.eu

Consistency of Trimmed Estimators of Scatter Under the t-distribution

Andrea Cerioli, Lucio Barabesi, Luis A. García-Escudero, and Agustín Mayo-Iscar

It is well known that trimmed estimators of multivariate scatter, such as the Minimum Covariance Determinant (MCD) estimator, are inconsistent unless an appropriate factor is applied to them in order to take the effect of trimming into account. This factor is widely recommended and applied when uncontaminated data are assumed to come from a multivariate Normal model (see, e.g., [4]). We address the problem of computing a consistency factor for the MCD estimator in a heavy-tail scenario, when uncontaminated data come from a multivariate Student's t-distribution. The multivariate t-distribution has a representation as an infinite mixture of Normals with scales depending on Gamma distribution. This representation allows estimation of the t-distribution parameters by using algorithms in the EM family. Additionally, the required consistency factor for trimmed estimators of multivariate scatter, such as the MCD, can be obtained through the corresponding consistency factors defined under the Normal model. We compare results from this mixture-based approach to analytical derivation of consistency factors that follows from the functional representation of the MCD [2]. We also consider implications for outlier detection [1] and robust clustering [3].

Keywords: consistency factor, em algorithm, mcd, robust distance

References

1. Cerioli, A.: Multivariate Outlier Detection With High-Breakdown Estimators. J. Am. Stat. Assoc. **105**, 147–156 (2010)
2. Croux, C., Haesbroeck, G.: Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. J. Multivar. Anal. **71**, 161–190 (1999)
3. Dotto, F., Farcomeni, A., Garcìa-Escudero, L.A., Mayo-Iscar, A.: A reweighting approach to robust clustering. Stat. Comp. **28**, 477–493 (2018)
4. Hubert, M., Debruyne, M., Rousseeuw, P.J.: Minimum covariance determinant and extensions. WIREs Comput Stat. **10**, e1421 (2018)

Andrea Cerioli
Department of Economics and Management, University of Parma,
e-mail: andrea.cerioli@unipr.it

Lucio Barabesi
Department of Economics and Statistics, University of Siena, e-mail: lucio.barabesi@unisi.it

Luis A. García-Escudero and Agustín Mayo-Iscar
Department of Statistics and OR, University of Valladolid,
e-mail: lagarcia@uva.es;agustin.mayo.iscar@uva.es

Robust Classification in High Dimensions Using Regularized Covariance Estimates

Valentin Todorov and Peter Filzmoser

High-dimensional highly correlated data exist in many application domains which requires the development of appropriate statistical methods. The classical classification methods like LDA and QDA become practically useless in such a setting because they will suffer from the singularity problem if the number of observed variables p exceeds the number of observations n. Numerous regularization techniques with the purpose to stabilize the classifier and achieve an improved classification performance have been developed and there exist several studies comparing various regularization techniques trying to facilitate the choice of a method. However, these methods are vulnerable to the presence of outlying observations (outliers) in the training data set which can influence the obtained classification rules and make the results unreliable. On the other hand, the high breakdown versions of discriminant analysis proposed in the literature, like [3] do not work or are not reliable in high dimensions. We propose to utilize the recently introduced regularized versions of the minimum covariance determinant (MCD) estimator - the regularized MCD (RMCD) estimator [2] and the minimum regularized covariance determinant (MRCD) estimator [1] to define the robust discriminant rules which will combine high robustness to outliers with applicability in high dimensions. The computations can be done with the R package **rrcov** available at CRAN. Simulated and real data examples show that the proposed methods perform better than the existing ones in a wide range of settings.

Keywords: regularization, high-dimensional classification, robust covariance estimation

References

1. Boudt, K., Rousseeuw, P.J., Vanduffel, S., Verdonck, T.: The minimum regularized covariance determinant estimator. Statistics and Computing **30**(1), 113–128 (2020).
2. Croux, C., Gelper, S., Haesbroeck, G.: Regularized minimum covariance determinant estimator. Technical report, Mimeo New York (2012).
3. Hubert, M., Van Driessen, K.: Fast and Robust Discriminant Analysis. Computational Statistics & Data Analysis **45**, 301–320 (2004).

Valentin Todorov
United Nations Industrial Development Organization (UNIDO), Vienna, Austria e-mail: `valentin@todorov.at`

Peter Filzmoser
Vienna University of Technology, Vienna, Austria e-mail: `p.filzmoser@tuwien.ac.at`

Symbolic Concordance and Discordance Illustrated on Data from an International Teaching and Learning Survey

Simona Korenjak-Černe, Barbara Japelj Pavešić, and Edwin Diday

A "similarity" as a "concordance" in data analysis represents mathematical modeling of the common words "similarity" and "concordance" used in our natural language. The similarity between a class c and a collection P of classes c' for a category x is high if the mean of the similarities between the frequency of x in c and the frequency of x in c' that varies in P is high. A class has high concordance with a given collection of classes for a category x if that category is frequent in that class and if, in addition, there are numerous classes in the given collection of classes for which the category x is also frequent. Similarity and concordance thus express two different kinds of knowledge. In this presentation, we introduce some measures of concordance and discordance between a class and a given collection of classes that fall within the framework of symbolic data analysis (SDA) [1].

We will illustrate the use of new measures on the real dataset from the international large-scale assessment PIRLS 2016 [2] that measured the achievement of students in classical reading and reading from digital devices in more than 50 countries. Online reading is becoming an extremely important skill for younger generations and research is needed to understand how it develops along with classical reading from paper. We will study distributions of high and low-achieving students in informational reading from paper and online reading, and compare these across countries and within teachers of classes of students inside a specific country. For example, if we consider the teacher as a class and his or her students as individuals, by examining their reading ability, we can measure the concordance or discordance between the teacher's student responses of a country and the other countries.

Keywords: symbolic data analysis, symbolic concordance, symbolic discordance

References

1. Diday, E.: Explanatory tools for machine learning in the symbolic data analysis framework. In: Diday, E., Guan, R., Saporta, G., Wang, H. (eds.) Advances in Data Science, pp. 3-30. ISTE-Wiley (2020)
2. PIRLS: Progress in Internationals Readings Literacys Study `https://timssandpirls.bc.edu/pirls2016`

Simona Korenjak-Černe
University of Ljubljana, School of Economics and Business, Slovenia, and Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia, e-mail: `simona.cerne@ef.uni-lj.si`

Barbara Japelj Pavešić
Educational Research Institute, Ljubljana, Slovenija, e-mail: `barbara.japelj@pei.si`

Edwin Diday
CEREMADE, University Paris-Dauphine, Paris, France, e-mail: `diday8@gmail.com`

A Clusterwise Regression Method for Distributional Data

Rosanna Verde, Antonio Balzanella, and Antonio Irpino

Abstract This paper deals with a cluster-wise regression method for distributional data. The set of objects to be clustered are described by distributional variables $\{Y, X_1, \ldots, X_p\}$, with Y the response variable and X_j's the predictors. Each object is represented by $p + 1$ probability functions, or empirical ones. Our proposal is based on a K-means clustering type-algorithm, where the centroid of the clusters are represented by linear regression models and the objects are assigned to the clusters according to minimum sum of squared errors. [1] and [2] proposed two regression models for distributional data based on a Non Linear Least Squared method and on the Wasserstein metric in a linear space. The constrain of non-negativity were imposed to guarantee the outcome is still a distributional variable. In consideration of the most recent developments in distributional data analysis (DDA), we introduce a transformation of the *qf*'s in quantile density functions [3], which allows to map density functions in an Hilbert space and overcome some challenge in DDA. Applications on synthetic and real data have corroborated the new method.

Keywords: symbolic data analysis, distributional data, quantile density functions

References

1. Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance, Advances in Data Analysis and Classification **9 (1)** 81-106 (2015)
2. Dias, S., Brito, P.: Linear regression model with histogram-valued variables, Statistical Analysis and Data Mining: The ASA Data Science Journal **8 (2)** 75-113 (2015)
3. Petersen, A., Muller, H.: Functional data analysis for density functions by transformation to a Hilbert space, Annals of Statistics **44 (1)** 183-218 (2016)

Rosanna Verde, Antonio Balzanella, Antonio Irpino
DMF, University of Campania, Italy
e-mail: \{rosanna.verde\},\{antonio.balzanella\},\{antonio.irpino\}@unicampania.it

The Use of Regression to Partition a Dataset of Interval Observations

Lynne Billard and Fei Liu

The use of regression modelling has an extensive history; likewise, clustering methodologies have existed for some time. In this work, we extend the dynamical partitioning concepts developed initially by Diday and Simon [1] combined with the k-means clustering approach of MacQueen [3], to a k-regression algorithm to enable clustering of interval-valued observations based on regression models. The usefulness of the algorithm is verified through some simulated data (as in the plots below) and applied to real data sets. More details can be found in [2].

Keywords: k-means, k-regression algorithm, dynamical partitioning

References

1. Diday, E., Simon, J. C.: Clustering analysis. In: Fu, K.S. (ed.) Digital Pattern Recognition, pp. 47-94. Springer, Berlin (1976)
2. Liu, F., Billard, L.: Partition of interval-valued observations using regression. Pattern Recognition (2022)
3. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: LeCam, L. M., Neyman, J. (eds.) Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley **1**, pp. 281-299 (1967)

Lynne Billard
University of Georgia, e-mail: lynne@stat.uga.edu

Fei Liu
Bank of America

Heterogeneous Random Forests

Ye-eun Kim and Hyunjoong Kim

Random forest(RF) is one of the most popular machine learning methods for classification problems. Two factors that affect the performance of RF are known to be the accuracy of individual trees and the diversity among trees. That is, the better the performance of each classifier and the more heterogeneous the individual classifiers, the better the RF performance. In this study, we propose a heterogeneous RF to increase the diversity of trees. The diversity was induced by intentionally creating a tree that is heterogeneous from the previous trees. Features used for splitting near the root node of the previous tree have lower weights when constructing the feature subspace of the next tree. Therefore, Features that were dominant in the previous tree are less likely to be used in the next tree and splitting features of root nodes becomes more diverse. As a result of comparing accuracy in several real data, Heterogeneous RF performed better than RF in data with dominant variables.

Keywords: ensemble, random forests, decision tree

References

1. Breiman, L. Random Forests. Machine Learning. **45**, 5–32 (2001)
2. Han, S., Kim, H. & Lee, YS. Double Random Forest. Machine Learning. **109**, 1569–1586 (2020)
3. Simon Bernar, Sebastien Adam & Laruent Heutte. Dynamic Random Forests. Pattern Recognition Letters, Elsevier. **33(12)**, 1580–1586 (2012)

Ye-eun Kim
Yonsei University, Address of Institute, e-mail: kyeun0628@yonsei.ac.kr

Hyunjoong Kim
Yonsei University, Address of Institute, e-mail: hkim@yonsei.ac.kr

Analysis of the Damage Rate Using Typhoon Information

Su Hoon Choi and Min Soo Kim

According to the Intergovernmental Panel on Climate Change, the scale and intensity of damage are increasing as well as the frequency of meteorological disasters due to global warming. Typhoon usually occur between July and October, so they are similar to the harvest of crops, causing a lot of damage to farmers. In order to minimize damage to farmers caused by meteorological disasters, South Korea has implemented crop insurance since 2001. This study aims to analyze typhoon and damage rate by using crop insurance data. Since crop insurance is measured objectively and fairly for accurate actual damage judgment, analysis using crop insurance data is expected to be highly reliable. It will identify and analyze the relationship between typhoon information and the damage rate, and further present the expected typhoon damage rate for future typhoon. Considering the characteristics of the analysis data, the zero-inflated beta regression will be used to analyze the damage rate. In addition, by using Random Forest and XGBoost, which are representative machine learning models, we intend to compare the prediction results between models. As a result of the prediction, the performance of machine learning models was better than zero-inflated beta regression. The results of this analysis are expected to be used not only to predict the expected damage rate for future typhoon but also to prepare measures to reduce damage to farmers.

Keywords: typhoon, crop insurance, zero-inflated beta regression, machine learning

References

1. Tang, B., Frye, H.A., Gelfand, A.E., Silander Jr, J.A.: Zero-inflated Beta distribution regression modeling. arXiv preprint arXiv:2112.07249 (2021)
2. Breiman, L.: Random forests. Machine learning, 45(1), 5-32 (2001)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794 (2016)

Su Hoon Choi
Department of Mathematics and Statistics, Chonnam National University, Gwangju, Korea
e-mail: magafand@naver.com

Min Soo Kim
Department of Statistics, Chonnam National University, Gwangju, Korea
e-mail: kimms@chonnam.ac.kr

Resampling, Relabeling, and Raking for Extremely Imbalanced Classification

Hae-Hwan Lee, Seunghwan Park, and Jongho Im

In this presentation, we consider the binary classification of extremely imbalanced data. Imbalanced data classification is often challengeable, especially for high-dimensional data, because unequal classes deteriorate classifier performance. Undersampling the majority class or oversampling the minority class are popular methods to construct balanced samples, facilitating classification performance improvement. However, many existing sampling methods cannot be easily extended to high-dimensional data and mixed data, including categorical variables, because they often require approximating the attribute distributions, which becomes another critical issue. To handle these issues, we propose a new sampling strategy employing resampling, relabeling, and raking procedures, such that the attribute values of the majority class are imputed for the values of the minority class in the construction of balanced samples. Our proposed algorithm is attractive in practice, considering that it does not require density estimation for synthetic data generation in oversampling and is not bothered by mixed-type variables. In addition, the proposed sampling strategy is robust to classifiers in the sense that classification performance is not sensitive to choosing the classifiers. Also the proposed method can be directly applied to one-class classification problem.

Keywords: calibration, mixed-type, one-class classification

Hae-Hwan Lee
Yonsei University, Seoul, South Korea, e-mail: 0210hwan@yonsei.ac.kr

Seunghwan Park
Kangwon University, Chuncheon, South Korea, e-mail: stat.shpark@kangwon.ac.kr

Jongho Im
Yonsei University, Seoul, South Korea, e-mail: ijh38@yonsei.ac.kr

Clustering Student Mobility Data in 3-way Networks

Vincenzo Giuseppe Genova, Giuseppe Giordano, Giancarlo Ragozini, and Maria Prosperina Vitale

The present contribution aims at introducing a network data reduction method for the analysis of 3-way networks [1] in which classes of nodes of different types are linked. The proposed approach enables simplifying a 3-way network into a weighted two-mode network by considering the statistical concept of joint dependence in a multiway contingency table. Starting from a real application on student mobility data in Italian universities [2], a 3-way network is defined, where provinces of residence, universities and educational programmes are considered as the three sets of nodes, and occurrences of student exchanges represent the set of links between them. The Infomap community detection algorithm [3] is then chosen for partitioning two-mode networks of students' cohorts to discover different network patterns.

Keywords: 3-way network, complex network, community detection, mobility data, tertiary education

References

1. Batagelj, V., Ferligoj, A., Doreian, P.: Indirect Blockmodeling of 3-Way Networks. In: Brito P., Cucumel G., Bertrand P., de Carvalho F. (eds) Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 151–159. Springer, Berlin, Heidelberg (2007)
2. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Geography of Italian student mobility: A network analysis approach. Socio Econ Plan Sci **73**, 100918 (2021)
3. Edler, D., Bohlin, L., Rosvall, M.: Mapping higher-order network flows in memory and multilayer networks with infomap. Algorithms, **10**, 112 (2017)

Vincenzo Giuseppe Genova
Department of Economics, Business, and Statistics, University of Palermo
e-mail: vincenzogiuseppe.genova@unipa.it

Giuseppe Giordano
Department of Political and Social Studies, University of Salerno, e-mail: ggiordano@unisa.it

Giancarlo Ragozini
Department of Political Science, Federico II University of Naples, e-mail: giragoz@unina.it

Maria Prosperina Vitale
Department of Political and Social Studies, University of Salerno, e-mail: mvitale@unisa.it

Multi-perspective Risky User Classification in Social Networks

Antonio Pellicani, Gianvito Pio, and Michelangelo Ceci

The widespread adoption of social media platforms opened up new ways to connect and engage in a globalized manner. However, it also led to the introduction of harmful addiction phenomena, and to the spread of cyberbullying and cyberterrorism activities. As a result, monitoring operations on the content published by users, as well as on their behavior, has become critical to ensure a correct and safe use of social medias. This monitoring process becomes very difficult in presence of *borderline* users, i.e., users who appear to act in safe way based on their posted content, but not according to other viewpoints (e.g., their relationships), and viceversa.

In this context, this abstract contributes towards an effective identification of risky users in social networks. Specifically, we propose a novel method that solves node classification tasks in social networks by exploiting the information conveyed by three different perspectives: the semantics of the textual content generated by users, the network of user relationships, and the users spatial closeness, derived from geo-tagging metadata associated with posted contents.

Existing approaches that consider multiple perspectives are mainly based on the injection of features identified from one perspective into the other [2], or are tailored for the analysis of the network structure and node attributes, without being able to capture the semantics of the generated content [1]. On the contrary, our method builds three models that exploit the peculiarities of each viewpoint, and learns a final model to fuse their contributions through a stacked generalization approach.

Our experiments on two variants of a real Twitter dataset showed that the proposed method outperforms 13 competitors based on one or more perspectives. This advantage is also clear on borderline users, confirming the applicability of our method in real-world social networks, which are potentially affected by noisy data.

Keywords: social network analysis, user risk identification, spatial analysis

References

1. Pio, G.: Multi-type clustering and classification from heterogeneous networks. Inf. Sci. **425**, 107–126 (2018)
2. Campbell, W.: Content+ context networks for user classification in twitter. NIPS 2014 Workshop (2014)

Antonio Pellicani, Gianvito Pio, Michelangelo Ceci
Dept. of Computer Science, University of Bari "Aldo Moro", Via Orabona, 4, 70125 Bari, Italy
e-mail: \{name.surname\}@uniba.it

Clustering and Blockmodeling Temporal Networks – Two Indirect Approaches

Vladimir Batagelj

Two approaches to clustering and blockmodeling of temporal networks are presented: the first is based on an adaptation of the clustering of symbolic data described by modal values and the second is based on clustering with relational constraints. Different options for describing a temporal block model are discussed.

Keywords: social networks, network analysis, blockmodeling, symbolic data analysis, clustering with relational constraints

Vladimir Batagelj
IMFM, Jadranska 19, 1000 Ljubljana, Slovenia,
IAM UP, Muzejski trg 2, 6000 Koper, Slovenia,
HSE, 11 Pokrovsky Bulvar, 101000 Moscow, Russian Federation
e-mail: vladimir.batagelj@fmf.uni-lj.si

Clustering Validation in Hierarchical Cluster Analysis: an Empirical Study

Osvaldo Silva, Áurea Sousa, and Helena Bacelar-Nicolau

The evaluation of clustering structures is a crucial step in cluster analysis. This study presents the main results of the hierarchical cluster analysis of variables concerning a real dataset in the context of Higher Education. The goal of this research is to find a typology of some relevant items taking into account both the homogeneity and the isolation of the clusters. Two similarity measures, namely the standard affinity coefficient and Spearman's correlation coefficient, were used, and combined with three probabilistic (*AVL*, *AVB* and *AV1*) aggregation criteria, from a parametric family in the scope of the *VL* (Validity Link) methodology [1]. The best partitions were selected based on some validation indices, namely the global *STAT* levels statistics and the measures P(I2, Σ) and γ [2], adapted to the case of similarity coefficients [3]. In order to evaluate the clusters and identify their most representative elements, the Mann and Whitney *U* statistics and the silhouette plot were also used.

Keywords: clustering validation, affinity coefficient, spearman correlation coefficient, *vl* methodology

References

1. Bacelar-Nicolau, H.: Contributions to the Study of Comparison Coefficients in Cluster Analysis (in Portuguese). Univ. Lisbon (1980)
2. Gordon, A.D.: Classification, 2nd Ed. Chapman & Hall, London (1999)
3. Silva, O., Bacelar-Nicolau, H., Nicolau, F. C., Sousa, Á.: Probabilistic approach for comparing partitions. In: Manca, R., McClean, S., Skiadas, C. H. (eds.) New Trends in Stochastic Modeling and Data Analysis, pp. 113-122. ISAST (International Society for the Advancement of Science and Technology), Athens (2015)

Osvaldo Silva
Universidade dos Açores and CICSNOVA.UAc, Rua da Mãe de Deus, 9500-321, Portugal
e-mail: osvaldo.dl.silva@uac.pt

Áurea Sousa
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, Portugal
e-mail: aurea.st.sousa@uac.pt

Helena Bacelar-Nicolau
Universidade de Lisboa (UL), Faculdade de Psicologia and Institute of Environmental Health (ISAMB/FM-UL), Portugal, e-mail: hbacelar@psicologia.ulisboa.pt

Divide and Conquer: a Clustering Method for Hierarchical and Nested Data Structures

Andrej Svetlošák, Miguel de Carvalho, Gabriel Martos Venturini, and Raffaella Calabrese

Joint clustering on data with nested or hierarchical structures can be challenging. Results obtained by similarity-based methods (i.e. via K-means and K-medoids) often do not reflect the structure of the data, while model-based clustering (i.e. via mixture models), as we show, likely to leads to the same number of components on each margin, i.e. same number of groups on each level of the hierarchy. We address these drawbacks of joint models by proposing a novel approach for cluster analysis—to which we refer to as *divide and conquer clustering*—that lies on the interface between model-based clustering and similarity-based clustering. The approach consists of three steps and provides interpretable cluster solutions while allowing differing number of components on the margins. We achieve this by first estimating the margins of each hierarchy level by recently introduced non-local prior mixtures, which have the advantage of treating the number of components as a model parameter. Secondly, we learn about a set of joint clusters (proto clusters) that are obtained via a Voronoi tessellation on the product space of the marginal component means. Finally, the final joint clusters are the Voronoi faces centred at local density maxima of the joint distribution. These are obtained by dividing up proto clusters with a density below a threshold between the remaining Voronoi faces. In this sense the high density areas *divide and conquer* low density regions. We analyse and compare the performance of our method with selected state of the art clustering methods. The results on both simulated data and real datasets suggest an on par or better performance than competing methods.

Keywords: cluster analysis, model-based clustering, similarity-based clustering, non-local priors, hierarchical and nested data structures

---

Andrej Svetlošák
Business School and School of Mathematics, The University of Edinburgh, 29 Buccleuch Pl, Edinburgh EH8 9JS, UK, e-mail: Andrej.Svetlosak@ed.ac.uk

Miguel de Carvalho
School of Mathematics, The University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK, e-mail: Miguel.deCarvalho@ed.ac.uk

Gabriel Martos Venturini
Departmento de Matemáticas y Estadística, Torcuato Di Tella University, Av. Figueroa Alcorta 7350 (C1428BCW) Ciudad de Buenos Aires, Argentina, e-mail: GMartos@utdt.edu

Raffaella Calabrese
Business School of The University of Edinburgh, 29 Buccleuch Pl, Edinburgh EH8 9JS, UK, e-mail: Raffaella.Calabrese@ed.ac.uk

Significance Mode Analysis (SigMA) for Hierarchical Structures

Sebastian Ratzenböck, Torsten Möller, Josefa E. Großschedl, João Alves, Immanuel M. Bomze, and Stefan Meingast

We present an innovative clustering method, Significance Mode Analysis (`SigMA`), to extract co-spatial and co-moving stellar populations from large-scale surveys such as ESA Gaia. The method studies the topological properties of the density field in the multidimensional phase space. The set of critical points in the density field gives rise to the cluster tree, a hierarchical structure in which leaves correspond to modes of the density function. Typically, however, non-parametric density estimation methods lead to an over-clustering of the input data. We propose an interpretable cluster tree pruning strategy by determining minimum energy paths between pairs of neighboring modes directly in the input space. We test for deviations from unimodality along these paths, which provides a measure of significance for each pair of clusters. We apply `SigMA` to Gaia data of the closest young stellar association to Earth, Scorpio-Centaurus (Sco-Cen), and find 48 co-moving clusters in Sco-Cen. These clusters are independently validated using astrophysical knowledge, to a certain extent, by their association with massive stars too bright for Gaia, both unknown to `SigMA`. Our findings suggest that Sco-Cen is more actively star-forming and dynamically richer than previously thought. This application demonstrates that `SigMA` allows for an accurate census of young populations, quantify their dynamics, and reconstruct the recent star formation history of the local Milky Way.

Keywords: mode seeking, cluster tree, stellar groups

Sebastian Ratzenböck
Data Science at Uni Vienna Research Network & Faculty of Computer Science, University of Vienna, Austria, e-mail: `sebastian.ratzenboeck@univie.ac.at`

Torsten Möller
Faculty of Computer Science & Data Science University of Vienna Research Network, Austria, e-mail: `torsten.moeller@univie.ac.at`

Josefa E. Großschedl
Department of Astrophysics, University of Vienna, Austria,
e-mail: `josefa.elisabeth.grossschedl@univie.ac.at`

João Alves
Department of Astrophysics & Data Science at University of Vienna Research Network, Austria,
e-mail: `joao.alves@univie.ac.at`

Immanuel M. Bomze
ISOR/VCOR & Data Science, University of Vienna Research Network, Austria,
e-mail: `immanuel.bomze@univie.ac.at`

Stefan Meingaset
Dep of Astrophysics, University of Vienna, Austria, e-mail: `stefan.meingast@univie.ac.at`

Kurtosis-based Projection Pursuit for Matrix-valued Data

Una Radojicic, Klaus Nordhausen, and Joni Virta

A classical problem in image processing is that of discriminatory feature extraction, where gray-scale images are naturally represented as matrices. We develop projection pursuit for data that admit a natural representation in matrix form, where another common data type admitting this representation is e.g. a univariate spatial data collected on a regular grid. For projection indices we propose extensions of the classical kurtosis and Mardia's multivariate kurtosis. The first index estimates projections for both sides of the matrices simultaneously, while the second index finds the two projections separately. Both indices are shown to recover the optimally separating projection for two-group Gaussian mixtures in the full absence of any label information. We further establish the strong consistency of the corresponding sample estimators. Simulations and a real data example on hand-written postal code data are used to demonstrate the method.

Keywords: discriminant analysis, matrix-variate gaussian mixture, rank-1 projection

References

1. Radojicic, U., Nordhausen K., and Virta J.: Kurtosis-based projection pursuit for matrix-valued data. arXiv preprint arXiv:2109.04167 (2021).

Una Radojicic
Vienna University of Technology, e-mail: una.radojicic@tuwien.ac.at

Klaus Nordhausen
University of Jyväskylä, e-mail: klaus.k.nordhausen@jya.fi

Joni Virta
University of Turku, e-mail: joni.virta@utu.fi

Comparison of Pixel Based Segmentation Methods in Papillary Thyroid US Images

Neslihan Gökmen İnan, İsmail Meşe, Düzgün Yıldırım, and Ozan Kocadağlı

Thyroid nodules are one of the endocrine diseases caused by abnormal growth of cells. Ultrasonography (US) is an efficient tool that is routinely used to identify these nodules. Thyroid nodule segmentation on US images is a valuable and, it has a great importance for the diagnosis of thyroid cancer. Despite US imaging is considered as the best option, the high incidence rate increases the burden of radiologists in terms of diagnosing the thyroid cancer cases at early stages and their levels [1]. In such a case, the contour and region-based segmentation methods have a potential to extract some important features called as biomarkers which allow radiologists to make more accurate diagnosis. In this context, this study aims to compare the contour and region-based segmentation methods with respect to the feature extraction performance. In this study, US images of 187 papillary carcinoma patients were analyzed. The image segmentation quality was evaluated with respect to dice coefficient measurement and ROC analysis results such as TN, FP, AUC, g-score, f-measure [2]. Also, the validity of the pixel-based methods were achieved with the extracted features obtained from the manual segmentation methods performed by expert radiologists. The analysis results showed that the automatic segmentation methods are superior performance to the manual ones according to the various statistical performance criteria.

Keywords: thyroid us image, pixel-based segmentation, papillary carcinoma

References

1. Chen, J., You, H., Li, K. : A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Computer methods and programs in biomedicine. **185**, 105329 (2020)
2. Koundal, D., Gupta, S., Singh, S. : Computer aided thyroid nodule detection system using medical ultrasound images. CBiomedical Signal Processing and Control. **40**, 117-130 (2018)

Neslihan Gökmen İnan
Koç University, Turkey, e-mail: ninan@ku.edu.tr

İsmail Meşe
Erenköy Hospital Mental and Nervous Diseases, Turkey, e-mail: ismail_mese@yahoo.com

Düzgün Yıldırım
Acibadem Kozyatagi Hospital, Turkey, e-mail: duzgun.yildirim@acibadem.com

Ozan Kocadağlı
Mimar Sinan Fine Arts University, Turkey, e-mail: ozan.kocadagli@msgsu.edu.tr

Bootstrapping Binary GEV Regressions for Massive Unbalanced Datasets

Michele La Rocca, Marcella Niglio, and Marialuisa Restaino

Research on rare events is constantly increasing over the years in many research areas. Examples include fraud detection, credit default prediction, bankruptcy prediction, customer/students churn predictions and accident occurrence. In all these cases, rare events data are usually defined as binary variables with fewer events (ones) than non-events (zeros). In other words, the degree of unbalance is more extreme in rare events than it is in the class of unbalanced data. However, both unbalanced and rare events data have been studied as statistical problems with possible applications in different fields such as biology, political science, engineering, economics and medicine. The unbalanced variables related to rare events are difficult to predict and explain, specially in high dimensional settings and in presence of massive datasets, where unbalancing might be even more critical.

The logistic model may not be appropriate for such data since it strongly underestimates the probability of rare events because the estimators tend to be biased towards the majority class, which is usually less critical. Moreover, as underlined in the literature, the bias of the maximum likelihood estimators of logistic regression parameters in small sample sizes could be amplified in a rare events context. Thus, in this framework, there is an increasing interest in using the quantile function of the GEV distribution as a link function to investigate the relationship between the binary response variable and a set of predictors. The main advantage of this approach is that thanks to its skewness, the GEV link function has an asymmetric behaviour. It approaches one slower than it approaches zero, handling nicely rare events.

This work aims to estimate the probability of success given a set of features by using a generalized extreme value regression model for binary data, also taking into account the effects on the response variable of class imbalance in categorical predictors. Confidence intervals and hypothesis testing are constructed by using bootstrap methods, specifically designed for massive datasets, in multiple testing perspectives. The performance of our proposed procedure is evaluated by Monte Carlo simulation studies and applications to real datasets.

Keywords: rare events data, GEV regression, bootstrap, multiple testing

---

Michele La Rocca
Department of Economics and Statistics - University of Salerno (Italy), e-mail: larocca@unisa.it

Marcella Niglio
Department of Economics and Statistics - University of Salerno (Italy), e-mail: mniglio@unisa.it

Marialuisa Restaino
Department of Economics and Statistics - University of Salerno (Italy),
e-mail: mlrestaino@unisa.it

Stochastic Collapsed Variational Inference for Structured Gaussian Process Regression Networks

Rui Meng, Herbert K. H. Lee, and Kristofer Bouchard

This paper presents an efficient variational inference framework for a family of structured Gaussian process regression network (SGPRN) models. We incorporate auxiliary inducing variables in latent functions and jointly treat both the distributions of the inducing variables and hyper-parameters as variational parameters. Then we take advantage of the collapsed representation of the model and propose structured variational distributions, which enables the decomposability of a tractable variational lower bound and leads to stochastic optimization. Our inference approach is able to model data in which outputs do not share a common input set, and with a computational complexity independent of the size of the inputs and outputs to easily handle datasets with missing values. Finally, we illustrate our approach on both synthetic and real data.

Keywords: stochastic optimization, gaussian process, variational inference, multivariate time series, time-varying correlation.

Rui Meng, Kristofer Bouchard
Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory
e-mail: rmeng@lbl.gov,kebouchard@lbl.gov

Herbert K. H. Lee
University of California, Santa Cruz e-mail: herbie@ucsc.edu

Covariate Selection Method in Propensity Score Model for the Quantile Treatment Effect Estimation

Takehiro Shoji, Jun Tsuchida, and Hiroshi Yadohisa

Estimation of a treatment effect, which is the impact of a treatment on an outcome, is important in some research areas, such as econometrics, social programs, and policies. Quantile treatment effects (QTE) are primarily used in econometrics because they can characterize the heterogeneous treatment effects on different points of an outcome distribution. For estimating QTE, Firpo [2] proposed an estimation method using propensity scores.
In estimating of treatment effects using propensity scores, selection of covariates to include propensity score model is an important issue, and it is known that it is better to include covariates that are relevant to the outcome [1]. For achieving this issues, Outcome Adaptive Lasso (OAL) was employed as a covariate selection method for propensity score models[2]. However, OAL assumes an average treatment effect estimation and not a quantile treatment effect estimation.
In this study, we propose a covariate selection method that includes propensity score models for quantile treatment effect estimation. Here, the central principle is changing the weight term from an outcome regression model to a quantile regression model. This allows for the selection of covariates related to an outcome at an interesting quantile corresponding to QTE. Through numerical experiments, we compare the proposed method's performance with that of the existing methods, such as OAL.

Keywords: propensity score, causal effect, quantile regression

References

1. Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T.: Variable selection for propensity score models. American journal of epidemiology. **163**(12), 1149–1156 (2006)
2. Firpo, S.: Efficient semiparametric estimation of quantile treatment effects. Econometrica. **75**(1), 259–276 (2007)
3. Shortreed, S. M., & Ertefaie, A.: Outcome-adaptive lasso: variable selection for causal inference. Biometrics, **73**(4), 1111–1122 (2017)

Takehiro Shoji
Graduate School of Culture and Information Science, Doshisha University, Tataramiyakodani 1-3, Kyotanabe City, Kyoto, Japan, e-mail: luckmt0107@gmail.com

Jun Tsuchida, Hiroshi Yadohisa
Department of Culture and Information Science, Doshisha University, Tataramiyakodani 1-3, Kyotanabe City, Kyoto, Japan

Are Attitudes Toward Immigration Changing in Europe? An Analysis Based on Latent Class IRT Models

Ewa Genge and Francesco Bartolucci

We analyze the changing attitudes toward immigration in EU host countries in the last few years (2010-2018) on the basis of the European Social Survey data. These data are collected by the administration of a questionnaire made of items concerning different aspects related to the immigration phenomenon. For this analysis, we rely on a latent class approach considering a variety of models that allow for: *(i)* multidimensionality; *(ii)* discreteness of the latent trait distribution; *(iii)* time-constant and time-varying covariates; and *(iv)* sample weights. Through these models we find latent classes of Europeans with similar levels of immigration acceptance and we study the effect of different socio-economic covariates on the probability of belonging to these classes for which we provide a specific interpretation. In this way we show which countries tend to be more or less positive toward immigration and we analyze the temporal dynamics of the phenomenon under study.

Keywords: discrete latent variables; european social survey; expectation-maximization algorithm; item response theory.

Ewa Genge
University of Economics in Katowice, Poland, e-mail: ewa.genge@ue.katowice.pl

Francesco Bartolucci
University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Visualization of IATA Regions in Air Transport Before and After the COVID-19 Pandemic

Tüzün Tolga İnan, Neslihan Gökmen İnan, Aylin Yaman Kocadağlı, and Ozan Kocadağlı

COVID-19 Pandemic has affected all transport modules including air passenger transportation (ATP) as an unprecedented global crisis. This study aims to visualize the position of International Air Transport Association (IATA) Regions considering some important key indicators such as gross domestic product, human development index, tourism arrival, national and international ATP numbers [1]. Also, this study deals with revealing out the negative impact of this global crisis on ATP during COVID-19 Pandemic between 2019-2021. In this study, to analyze ATP, three common airline metrics were gathered to figure out significant factors that lead to similarities between regions in terms of global crisis's negative impact for each year. Factor analysis (FA) and multidimensional scaling (MDS) were applied to IATA dataset [2]. MDS has brought out some substantial inferences. For instance, Europe, North and Latin America have similarities in positively, whereas Africa, Asia Pacific and Middle East show these similarities negatively since they are located close to each other. The location of regions has changed due to COVID-19 Pandemic compared to 2019 in MDS. Thus, ATP recovery is better in the Middle East compared to Africa and the Asia Pacific; however, this recovery circulation seems far from being adequate when compared to others. To sum up, these findings may help aviators to manage the strategic perspective of ATP more professionally.

Keywords: covid-19 pandemic, air passenger transportation, strategic perspective

References

1. Scholl, W., Schermuly, C.C.: The impact of culture on corruption, gross domestic product, and human development. Journal of Business Ethics **163(1)**, 171–189 (2020)
2. Nyoja, E.T., Ragab, M.R.: Economic Impacts of Public Air Transport Investment: A Case Study of Egypt. Sustainability **14(5)**, 2651 (2022)

Tüzün Tolga İnan
Bahçeşehir University, Turkey, e-mail: tuzuntolga.inan@sad.bau.edu.tr

Neslihan Gökmen İnan
Koç University, Turkey, e-mail: ninan@ku.edu.tr

Aylin Yaman Kocadağlı
İstanbul University, Turkey, e-mail: yaman@istanbul.edu.tr

Ozan Kocadağlı
Mimar Sinan Fine Arts University, Turkey, e-mail: ozan.kocadagli@msgsu.edu.tr

Political and Religion Attitudes in Greece: Behavioral Discourses

Georgia Panagiotidou and Theodore Chadjipadelis

The research presented in this paper attempts to explore the relationship between religious and political attitudes. More specifically we investigate how religious behavior, in terms of belief intensity and practice frequency, is related to specific patterns of political behavior such as ideology, understanding democracy and his set of moral values. The analysis is based on the use of multivariable methods and more specifically Hierarchical Cluster Analysis and Multiple Corre-spondence Analysis in two steps. The findings are based on a survey implemented in 2019 on a sample of 506 respondents in the wider area of Thessaloniki, Greece. The aim of the research is to highlight the role of people's religious practice intensity in shaping their political views by displaying the profiles resulting from the analysis and linking individual religious and political characteristics as measured with various variables. The final output of the analysis is a map where all variable categories are visualized, bringing forward models of political behavior as associated together with other factors such as religion, moral values and democratic attitudes.

Keywords: political behavior, religion, democracy, multivariate methods, data analysis

References

1. Greenacre, M.: Correspondence Analysis in Practice. Chapman and Hall/CRC Press, Boca Raton (2007)
2. Marangudakis, M. Chadjipadelis, T.: The Greek Crisis and its Cultural Origins. Palgrave-Macmillan, New York (2019)
3. Mayer, N.: Les modeles explicatifs du vote. L'Harmatan, Paris (1997)
4. Michelat, G., Simon, M.: Classe, Religion et Comportement Politique. PFNSP-Editions Sociales, Paris (1977)
5. Panagiotidou, G., Chadjipadelis, T.: First-time Voters in Greece: Views and Attitudes of Youth on Europe and Democracy. In Theodore Chadjipadelis, Berthold Lausen, Angelos Markos, Tae Rim Lee, Angela Montanari and Rebecca Nugent (Eds), Studies in Classification, Data Analysis and Knowledge Organization, 415-429, Springer (2020)
6. Papadimitriou, G., Florou, G.: Contribution of the Euclidean and chi-square metrics to determining the most ideal clustering in ascending hierarchy (in Greek). In An-nals in Honor of Professor I. Liakis, 546-581. University of Macedonia, Thessaloniki (1996)
7. Rose, R.: Electoral Behavior: a comparative Handbook. Free Press, New York (1974)

Georgia Panagiotidou · Theodore Chadjipadelis
Aristotle University of Thessaloniki, e-mail: {gvpanag,chadji}@polsci.auth.gr

Functional Data Representation with Merge Trees

Matteo Pegoraro and Piercesare Secchi

Topological Data Analysis is a branch of data analysis aiming at representing data by means of topological information. Such representations are very different from classical statistical ones and posses interesting properties that can be used to tackle data analysis problems with a different perspective. We will represent functions by means of objects called merge trees, and with their properties we will look at the problem of alignment and smoothing of functional data within a benchmark case study.

Keywords: functional data analysis, topological data analysis, merge trees, functional registration, smoothing

Matteo Pegoraro
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: matteo.pegoraro@polimi.it

Piercesare Secchi
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, e-mail: piercesare.secchi@polimi.it

Elastic Regression for Irregularly Sampled Curves in $\mathbb{R}^d$

Lisa Steyer, Almond Stöcker, and Sonja Greven

We propose regression models for curve-valued responses in two or more dimensions, where only the image but not the parametrisation of the curves is of interest. Examples of such data are handwritten letters, movement paths or outlines of objects. In the square-root-velocity framework [1], a parametrisation invariant distance for curves is obtained as the quotient space metric with respect to the action of re-parametrisation, which is by isometries. With this special case in mind, we discuss the generalisation of 'linear' regression to quotient spaces more generally, before illustrating the usefulness of our approach for curves modulo re-parametrisation. We test this model in simulations and apply it to human hippocampi data, obtained from MRI scans [2]. Here we model how the shape of the hippocampus is related to age and Alzheimer's disease. We address the issue of irregularly sampled curves by using splines for modelling smooth predicted curves.

Keywords: elastic regression, sparse functional data, square-root-velocity framework, warping

References

1. Srivastava, A. and Klassen, E.P.: Functional and Shape Data Analysis. In: Springer Series in Statistics. Springer New York (2016)
2. Petersen, R.C. et al.: Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. In: Neurology, 74(3):201–209, 2010.

Lisa Steyer
Humboldt University of Berlin, Unter den Linden 6, 10117 Berlin,
e-mail: lisa.steyer@hu-berlin.de

Almond Stöcker
Humboldt University of Berlin, e-mail: almond.stoecker@hu-berlin.de

Sonja Greven
Humboldt University of Berlin, e-mail: sonja.greven@hu-berlin.de

Misalignment of Spectral Data: Constrained Optimization in a Functional Data Analysis Framework

Francesca Di Salvo, Delia Francesca Chillura Martino, and Gabriella Chirco

Across several branches of sciences, a large number of applications involves data represented as functions and curves, for which functional data analysis can play a central role in solving a variety of problem formulations. With some thecnologies, the obtained data are spectra containing a vast amount of information concerning the composition of a sample: in order to infer the chemical composition of the materials from spectra, functional data analysis offers a valuable mean for characterizing the spectral response through identification of peaks position and intensity. The collection of data from different measurement may exhibit similar peak pattern but display misalignment in their peaks. In general, the multiple alignment is crucial in the subsequent analysis; the method proposed faces with the challenge of random shifts in the peaks and implements constraints in a proper objective function to optimize the alignment. The constraints are based on a priori information that is formalized in the choice of a set of peaks across functions. Spectrum data from X-ray Fluorescence (XRF) and Total Reflectance-Fourier Transform Infra-Red (TR-FTIR) spectroscopies are considered to illustrate the approach and to provide useful comparison with other approaches.

Keywords: multple alignment, functional data analysis, constrained registration

References

1. Houhou, R., Rösch, P., Popp, J., Bocklitz, T.: Comparison of functional and discrete data analysis regimes for Raman spectra. Anal. and Bioanal. Chem.**413**, 5633–5644, (2021)
2. Marron,J. S., Ramsay, J. O., Sangalli L. M., Srivastava, A.: Functional Data Analysis of Amplitude and Phase Variation. Stat. Sci. **30**, No. 4, 468–484, (2015)
3. Srivastava,A., Klassen, E.P.: Functional and Shape Data Analysis. Springer Series in Statistics, Springer-Verlag, New York (2016)

Francesca Di Salvo
Department of Agricultural Food and Forest Sciences - SAAF, Università degli Studi di Palermo, Viale deljle Scienze ed.4, Palermo I-90128, Italy e-mail: francesca.disalvo@unipa.it

Delia Francesca Chillura Martino, Gabriella Chirco
Dipartimento Scienze e Tecnologie Biologiche, Chimiche e Farmaceutiche – STEBICEF, Università degli Studi di Palermo, Viale delle Scienze ed.17, Palermo I-90128, Italy,
e-mail: delia.chilluramartino@unipa.it, gabriella.chirco@community.unipa.it

Model Based Clustering of Functional Data with Mild Outliers

Cristina Anton and Iain Smith

We propose a procedure, called CFunHDDC, for clustering functional data with mild outliers which combines two existing clustering methods: the functional high dimensional data clustering (FunHDDC) [1] and the contaminated normal mixture (CNmixt) [2] method for multivariate data. We adapt the FunHDDC approach to data with mild outliers by considering a mixture of multivariate contaminated normal distributions. To fit the functional data in group-specific functional subspaces we extend the parsimonious models considered in FunHDDC, and we estimate the model parameters using an expectation-conditional maximization algorithm (ECM). The performance of the proposed method is illustrated for simulated and real-world functional data, and CFunHDDC outperforms FunHDDC when applied to functional data with outliers.

Keywords: functional data, model-based clustering, contaminated normal distributions, em algorithm

References

1. Bouveyron, C., Jacques, J: Model-based clustering of time series in group-specific functional subspaces. Adv. Data. Anal. Classif. **5**(4), 281–300 (2011)
2. Punzo, A., McNicholas, P.D.: Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J., **58**, 1506–1537 (2016)

Cristina Anton
MacEwan University, 10700 – 104 Avenue Edmonton, AB, T5J 4S2, Canada
e-mail: popescuc@macewan.ca

Iain Smith
MacEwan University, 10700 – 104 Avenue Edmonton, AB, T5J 4S2,Canada
e-mail: smithi23@mymacewan.ca

Old and New Constraints in Model Based Clustering

Luis A. García-Escudero, Agustín Mayo-Iscar, Gianluca Morelli, and Marco Riani

Model-based approaches to cluster analysis and mixture modeling often involve maximizing classification and mixture likelihoods. Without appropriate constrains on the scatter matrices of the components, these maximizations result in ill-posed problems. Moreover, without constrains, non-interesting or "spurious" clusters are often detected by the EM and CEM algorithms traditionally used for the maximization of the likelihood criteria. A useful approach to avoid spurious solutions is to restrict relative components scatter by a prespecified tuning constant. Recently new methodologies for constrained parsimonious model-based clustering have been introduced which include the 14 parsimonious models that are often applied in model-based clustering when assuming normal components as limit cases. In this paper we initially review the traditional approaches and illustrate through an example the benefits of the adoption of the new constraints.

Keywords: model based clustering, mixture modelling, constraints

L.A. García-Escudero
Department of Statistics and Operational Research and IMUVA, University of Valladolid
e-mail: lagarcia@eio.uva.es

A. Mayo-Iscar
Department of Statistics and Operational Research and IMUVA, University of Valladolid
e-mail: agustinm@eio.uva.es

G. Morelli
Department of Economics and Management and Interdepartmental Centre of Robust Statistics, University of Parma, e-mail: gianluca.morelli@unipr.it

M. Riani
Department of Economics and Management and Interdepartmental Centre of Robust Statistics, University of Parma, e-mail: mriani@unipr.it

Model Based Clustering and Outlier Detection with Missing Data

Cristina Tortora, Hung Tong, and Louis Tran

Cluster analysis is a data analysis technique that aims to produce smaller groups of similar observations in a data set. In model-based clustering, the population is assumed to be a convex combination of sub-populations, each of which is modeled by a probability distribution. When the data are characterized by outliers the multivariate Student-t (T) and the contaminated normal distribution (CN) provide robust parameter estimates and therefore are more suitable choices compared to Gaussian Mixture models. Recently, the T and CN distributions have been extended to accommodate different tail behaviors across principal components, the models are referred to as multiple scaled distributions, i.e., MST and MSCN respectively. The mixture of CN has the advantage of automatically detecting outliers while the MSCN distribution, has the advantage of directional robust parameter estimates and outlier detection. The term "directional" implies that the parameter estimation and outlier detection procedures work separately for each principal component. Some practical limitations of the mentioned models are that they require the number of clusters to be known and the data set to be complete. This work has overcome the two mentioned limitations providing a study of indices to select the number of clusters and presenting recent extensions of the CN and MSCN mixtures to cluster data that contain values missing at random. All the discussed techniques are available in two convenient R packages MSclust and MixtureMissing.

Keywords: model based clustering, outlier detection, missing values

References

1. Tong H., Tortora. C. Model-based clustering and outlier detection with missing data. Advances in data analysis and classification, 1-26 2022
2. Tran L. and Tortora C. How Many Clusters Are Best? Investigating Model Selection in Robust Clustering. In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 1159-1180 2021
3. Tong H., Tortora. C. MixtureMissing: Robust Model-Based Clustering for Data Sets with Missing Values at Random. R package v. 1.0.2. 2022
4. Tortora. C., Punzo A., Tran L. MSclust: Multiple-Scaled Clustering. R package v. 1.0. 2022

Cristina Tortora and Louis Tran
San José State University, San José, California, USA, e-mail: cristina.tortora@sjsu.edu

Hung Tong
The University of Alabama,Tuscaloosa, Alabama, USA e-mail: hungtongmx@gmail.com

How to Mitigate the Effect of Outliers on Balancing Technique

Rasool Taban, Maria do Rosário Oliveira, and Claudia Nunes Philippart

Imbalanced data brings additional difficulties to the analysis of the data since the distribution of the number of observations across the known classes is skewed. The skewness implies that the number of observations in the Minority class is drastically smaller than the number of observations in the Majority class. This is a problem since, typically, the Minority class is the interesting and relevant class. Many real-world applications face this issue due to their natural characteristics, such as fraud detection, rare disease detection, etc.
Balancing techniques are a common strategy to overcome imbalanced data problems, but the presence of outliers may lead to bias and poor results, especially when the outliers are located in the Minority class and we use classical methods.
In this work, first, we illustrate the negative effect of outliers on the performance of classical balancing techniques. Next, we propose a robust balancing technique to mitigate the effect of outliers - named RM-SMOTE – which combines the idea of SMOTE with robust Mahalanobis distance. We propose to automatically down weight atypical Minority class observations so that they have a low chance of being selected in the resampling step.
The performance of the RM-SMOTE is evaluated using simulated data with different levels of contamination, and benchmark imbalanced datasets. The results indicate the superiority of RM-SMOTE when handling different proportions of outliers. In cases where the observations are not linearly separable, RM-SMOTE superiority is even more evident.

Keywords: imbalanced data, balancing techniques, robust mahalanobis distance, over-sampling, smote

Rasool Taban
CEMAT and Department of Mathematics, Instituto Superior Técnico, Lisbon, Portugal
e-mail: rasooltaban@gmail.com

Maria do Rosário Oliveira
CEMAT and Department of Mathematics, Instituto Superior Técnico, Lisbon, Portugal
e-mail: rosario.oliveira@tecnico.ulisboa.pt

Claudia Nunes Philippart
CEMAT and Department of Mathematics, Instituto Superior Técnico, Lisbon, Portugal
e-mail: cnunes@math.tecnico.ulisboa.pt

Outliers Detection in Functional Data

Amovin-Assagba Martial, Gannaz Irène, and Jacques Julien

The modern technologies ease the collection of massive data at high frequency. From a statistical point of view, these data can be considered as `functional data`: discrete observations of random functions. One of the key problems in functional data analysis, is the `detection of outliers`. For this purpose, we propose a robust method based on `contaminated Gaussian mixture models` [1]. This model allows both to group and to detect outliers in multivariate `functional data`. A mixture of multivariate contaminated Gaussian distributions [2] is a Gaussian mixture where each cluster has two components: one, with a large prior probability, represents normal observations, and the other, with a small prior probability, represents outliers. Dimension reduction methods based on [3], are used to introduce parsimony into the model. An ECM (Expectation-Conditional Maximization) algorithm is proposed for model inference and the choice of hyper-parameters is addressed through model selection. The model performs efficiently on simulated data. It also helps to correctly detect outliers in the industrial data sets which motivated this work.

Keywords: outlier detection, contaminated gaussian mixture model, functional data, model-based clustering, em algorithm

References

1. Amovin-Assagba, M., Gannaz, I., & Jacques, J.: Outlier detection in multivariate functional data through a contaminated mixture model. arXiv preprint arXiv:2106.07222 (2021)
2. Punzo, A., & McNicholas, P. D. : Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J., **58**, 1506-1537 (2016)
3. Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. : Clustering multivariate functional data in group-specific functional subspaces. Comput. Stat., **35**, 1101-1131 (2020)

Amovin-Assagba Martial
Arpege Master K, 15 rue du dauphiné, 69800, Saint-Priest, France / Univ Lyon, Univ Lyon 2, ERIC UR3083, Lyon, France, e-mail: `martial.amovin@masterk.com`

Gannaz Irène
Univ Lyon, INSA Lyon, UJM, UCBL, ECL, ICJ, UMR5208, Villeurbanne, 69621, France
e-mail: `irene.gannaz@insa-lyon.fr`

Jacques Julien
Univ Lyon, Univ Lyon 2, ERIC UR3083, Lyon, France,
e-mail: `julien.jacques@univ-lyon2.fr`

Robustified Elastic Net Estimator for Multinomial Regression

Fatma Sevinç Kurnaz and Peter Filzmoser

The elastic net estimator has been proposed in particular for high-dimensional low sample size data sets [5], and it has been extended to generalized linear regression models [1]. A fully robust version of the elastic net estimator has been introduced for linear and logistic regression by [3]. This work is extended to the setting of robust multinomial regression. Robustness is achieved by trimming the negative log-likelihood function, and by introducing group-wise weights according to the outlyingness of the observations. The procedure is implemented in the R package *enetLTS* [4], using internally the R package *glmnet* [2]. Simulation studies and real data examples are conducted to show the performance in comparison to the classical, non-robust counterpart for multinomial regression.

The work was supported by grant TUBITAK 2219 from the Scientific and Technological Research Council of Turkey.

Keywords: elastic net penalty, multinomial regression, robustness, sparsity

References

1. Friedman, J., Hastie, T. and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, **33**, 1–22, 2010.
2. Friedman, J., Hastie, T. and Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J. and Yang, J. 2021: *glmnet*: Lasso and Elastic-Net Regularized Generalized Linear Models, R Foundation for Statistical Computing, Vienna, Austria. R package version 4.1–3, https://CRAN.R-project.org/package=glmnet
3. Kurnaz, F.S, Hoffmann, I. and Filzmoser, P.: Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemometrics and Intelligent Laboratory Systems, **172**, 211–222 (2018).
4. Kurnaz, F.S, Hoffmann, I. and Filzmoser, P.: *enetLTS*: Robust and sparse estimation methods for high-dimensional linear and logistic regression, R Foundation for Statistical Computing, Vienna, Austria. R package, https://CRAN.R-project.org/package=enetLTS
5. Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, Journal of Royal Statistics Society Series B, **67**, 301–320 (2005).

Fatma Sevinç Kurnaz
Department of Statistics, Yildiz Technical University, 34220 Istanbul, Turkey,
e-mail: fskurnaz@yildiz.edu.tr

Peter Filzmoser
TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria, e-mail: peter.filzmoser@tuwien.ac.at

Optimized Symbolic Correspondence Analysis for Multi-valued Variables

Jorge Arce Garro and Oldemar Rodríguez Rojas

In this paper, we propose an Optimized Correspondence Factorial Analysis (OCFA) method to analyze a data table with set-valued symbolic variables. This is an extension of Symbolic Correspondence Factorial Analysis (SCFA). OCFA is a combination between Symbolic Correspondence Factorial Analysis, based on an interval contingency data table and integer optimization. The idea is to choose the best matrix of integer values inside the interval contingency data table. We are interested in studying two different objective functions: the first one search to minimize the distance between projections and the original points, while the second one search to maximize the explained variance. To solve these problems, we generalize the concepts of row and column profiles to interval row and interval column profiles, respectively. Further, we propose two theorems to find the coordinates of the interval contingency table in the factorial axes. All of the methods proposed in this paper can be executed in the RSDA package, developed in R that can be downloaded from CRAN.

Keywords: symbolic data analysis, correspondence analysis, multi-valued variables, interval contingency table

References

1. Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* (United Kingdom: John Wiley & Sons Ltd)
2. Rodriguez, O. (2021). *RSDA: R to Symbolic Data Analysis*. R package version 3.0.9
3. Takagia, I. and Yadohisab, H. (2011). Correspondence analysis for symbolic contingency tables based on interval algebra. *Procedia Computer Science* 6, 352–357. 10.1016/j.procs.2011.08.065

Jorge Arce Garro
School of Mathematics, National University of Costa Rica, Costa Rica
e-mail: jorge.arce.garro@una.ac.cr

Oldemar Rodríguez Rojas
School of Mathematics, Research Center in Pure and Applied Mathematics (CIMPA), University of Costa Rica, Costa Rica, e-mail: oldemar.rodriguez@ucr.ac.cr

Symbolic t-SNE and UMAP Methods for Interval Type Variables.

Oldemar Rodríguez Rojas

UMAP (Uniform Manifold Approximation and Projection) is a very new method for dimension reduction. UMAP method improve t-SNE (t-Distributed Stochastic Neighbor Embedding) method for data visualization and dimensionality reduction. The great advantage of UMAP is that it preserves better than t-SNE the global structure with superior run time performance. The foregoing makes UMAP an ideal method to be applied to the hyper-rectangles that are in the rows of the symbolic data table with interval-type variables, since UMAP compresses the structure inside each hyper-rectangle very well and at the same time better preserves the global structure of the clusters generated by each hyper-rectangle. This paper presents an adapted version of the t-SNE and UMAP methods for interval type variables. In addition, R and Python codes for both generalizations are presented.

Keywords: symbolic data analysis, t-SNE, umap, interval variables.

References

1. Arce, J. and Rodríguez, O. (2019). Optimized dimensionality reduction methods for interval-valued variables and their application to facial recognition. *Entropy 2019* https://doi.org/10.3390/e21101016
2. Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* (United Kingdom: John Wiley & Sons Ltd)
3. Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997). Extension de l'analyse en com-posantes principales à des données de type intervalle. *Statistique Appliquée* XLV 3, 5–24
4. Douzal-Chouakria, A., Billard, L., Diday, E., and Schektman, Y. (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining* XLV 4, 229–246. 10.1002/sam
5. Laurens Van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9,86,2579-2605, http://jmlr.org/papers/v9/vandermaaten08a.html
6. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap* 6, 352–357. http://arxiv.org/abs/1802.03426
7. Rodríguez, O. (2007). Correspondence analysis for symbolic multi-valued variables. *Carme 2007*, Rotterdam, The Netherlands. http://www.carme-n.org/carme2007/.
8. Rodriguez, O. (2021). *RSDA: R to Symbolic Data Analysis*. R package version 3.0.9

Oldemar Rodríguez Rojas
School of Mathematics, Research Center in Pure and Applied Mathematics (CIMPA), University of Costa Rica, Costa Rica e-mail: oldemar.rodriguez@ucr.ac.cr.

Two-stage Principal Component Analysis on Interval-valued Data Using Patterned Covariance Structure

Anuradha Roy

A new approach is developed for facial recognition using principal component analysis of interval-valued data. We exploit patterned covariance structures in doing so and we accomplish this in two stages: first, we get eigenblocks and eigenmatrices of the patterned variance-covariance matrix, and then we analyze these eigenblocks and the corresponding principal vectors together in some appropriate way to get the principal components of the interval-valued data. We apply our method to the face recognition data in [2]. We take care of the three sequences of each face by using structured covariance matrices and answer the question whether three sequences belong to the same face or not. Face sequence recognition or classification is an important problem as face might slightly change due to several reasons. Results illustrating the accuracy and appropriateness of the new method over the existing methods are presented.

Keywords: interval-valued data, patterned covariance structures, eigenblocks and eigenmatrices, principal vectors

References

1. Douzal-Chouakria, A., Billard, L. and Diday E.: Principal component analysis for interval-valued observations. Stat. Anal. Data Min.: The ASA Data Science Journal, **4(2)**, 229-246 (2011)

Anuradha Roy
The University of Texas at San Antonio, Department of Management Science and Statistics
e-mail: Anuradha.Roy@utsa.edu

Detection of the Biliary Atresia Using Deep Convolutional Neural Networks Based on Statistical Learning Weights via Optimal Similarity and Resampling Methods

Kuniyoshi Hayashi, Eri Hoshino, Mitsuyoshi Suzuki, Erika Nakanishi, Kotomi Sakai, and Masayuki Obatake

Recently, artificial intelligence methods have been applied in several fields, and their usefulness is attracting attention. Neural networks are representative online models for prediction and discrimination. Many online methods require large training data to attain sufficient convergence. Thus, online models may not converge effectively for low and noisy training datasets. For such cases, to realize effective learning convergence in online models, we introduce statistical insights into an existing method to set the initial weights of deep convolutional neural networks. Using an optimal similarity and resampling method, we proposed an initial weight configuration approach for neural networks. For a practice example, identification of biliary atresia (a rare disease), we verified the usefulness of the proposed method by comparing existing methods that also set initial weights of neural networks.

Keywords: auc, bootstrap method, sensitivity and specificity, projection matrix

Kuniyoshi Hayashi
Graduate School of Public Health, St. Luke's International University, 3-6 Tsukiji, Chuo-ku, Tokyo, Japan, 104-0045, e-mail: khayashi@slcn.ac.jp

Eri Hoshino · Kotomi Sakai
Research Organization of Science and Technology, Ritsumeikan University, 90-94 Chudoji Awatacho, Shimogyo Ward, Kyoto, Japan, 600-8815,
e-mail: erihoshino119@gmail.com;koto.sakai1227@gmail.com

Mitsuyoshi Suzuki
Department of Pediatrics, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan, 113-8421, e-mail: msuzuki@juntendo.ac.jp

Erika Nakanishi
Department of Palliative Nursing, Health Sciences, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Japan, 980-8575
e-mail: nakanishi.erika.q3@dc.tohoku.ac.jp

Masayuki Obatake
Department of Pediatric Surgery, Kochi Medical School, 185-1 Kohasu, Oko-cho, Nankoku-shi, Kochi, Japan, 783-8505, e-mail: mobatake@kochi-u.ac.jp

Variational Autoencoder with Gamma Mixture for Clustering Right-skewed Data

Jinwon Heo and Jangsun Baek

Generative models such as generative adversarial network(GAN), autoencoder(AE), and variational autoencoder(VAE) enable model-based clustering to be undertaken because they can learn and extract significant features from data. Among them, variational autoencoder with deep embedding(VaDE) [2, 3] is an unsupervised clustering method proposed within the VAE framework by assuming Gaussian distribution for both the marginal distribution of the latent feature vector and the conditional distribution of data given the latent vector. Several analyses of many real microarray datasets have suggested that the empirical distribution of gene expression levels is approximately right-skewed like log-normal with some extreme values depending on the biological samples under investigation. Therefore, the above approach is sensitive to both non-normality of the data and extreme expression levels. We propose a new VAE approach based on gamma mixture that efficiently fits data with right-skewed distribution. We derive the evidence lower bound (ELBO) and optimize the ELBO using the reparameterization trick for gamma distribution and Stochastic Gradient Variational Bayes estimator. The proposed method is applied to some high-dimensional real gene expression datasets and single-cell RNA-seq data with small sample sizes and shows its better performance over the existing generative models including statistical model-based method such as mixtures of common t-factor analyzers [3].

Keywords: clustering, variational autoencoder, gamma distribution

References

1. Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148. (2016)
2. Yang, L., Fan, W., Bouguila, N.: Clustering analysis via deep generative models with mixture models. IEEE Transactions on Neural Networks and Learning Systems. **33(1)**, 340-350 (2020)
3. Baek, J., McLachlan, G. J.: Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics. **27(9)**, 1269-1276 (2011)

Jinwon Heo
Department of Mathematics and Statistics, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, e-mail: 206196@jnu.ac.kr

Jangsun Baek
Department of Mathematics and Statistics, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, e-mail: jbaek@jnu.ac.kr

An Efficient Way to Identify Inliers via Inlier-memorization Effect of Deep Generative Models

Dongha Kim, Jaesung Hwang, and Yongdai Kim

Identifying whether a given sample is an outlier or not is a significant issue in various real-world domains. Many trials have developed outlier detection methods, but they mainly presumed no outliers in the training data set. This study considers a more general situation where training data contains some outliers, and any information about inliers and outliers is not given. We propose a powerful and efficient learning framework to identify inliers in a training data set using deep neural networks. We start with a new observation, called the inlier-memorization effect, that when we train a deep generative model with data contaminated with outliers, the model first memorizes inliers before outliers. Exploiting this finding, we develop a new method called Outlier Detection via the Inlier-Memorization effect (ODIM). The ODIM only requires a few updates; thus, it is fast and efficient. We also provide a data-adaptive strategy to find the optimal number of updates, which makes the ODIM applied to real domains at ease. We empirically demonstrate that our method can refine inliers successfully in both tabular and image data sets.

Keywords: unsupervised anomaly detection, deep generative models, inlier-memorization effect

Dongha Kim
School of Mathematics, Statistics, and Data Science, Data Science Center, Sungshin Women's University, 2, 34 da-gil, Bomun-ro, Seongbuk-gu, Seoul, Republic of Korea,
e-mail: dongha0718@sungshin.ac.kr

Jaesung Hwang
SK Telecom, 264, Pangyo-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea,
e-mail: postechiminuru@gmail.com

Yongdai Kim
Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea,
e-mail: ydkim903@snu.ac.kr

Three-way Spectral Clustering

Cinzia Di Nuzzo and Salvatore Ingrassia

In this paper, we present a spectral clustering approach for clustering *three-way data*. Three-way data concerns data characterized by three modes: n units, p variables, and t different occasions. In other words, three-way data contain a $t \times p$ observed matrix for each statistical observation. The units generated by simultaneous observation of variables in different contexts are usually structured as three-way data, so each unit is basically represented as a matrix. The spectral clustering application to three-way data can be a powerful tool for unsupervised classification. Here, one example on real three-way data have been presented showing that spectral clustering method is a competitive method to cluster this type of data.

Keywords: spectral clustering, kernel function, three-way data

Cinzia Di Nuzzo
Department of Economics and Business, University of Catania, Piazza Università, 2, 95131 Catania,
e-mail: cinzia.dinuzzo@phd.unict.it

Salvatore Ingrassia
Department of Economics and Business, University of Catania, Piazza Università, 2, 95131 Catania,
e-mail: s.ingrassia@unict.it

Fuzzy Clustering by Hyperbolic Smoothing

David Masís, Esteban Segura, Javier Trejos, and Adilson Xavier

We propose a novel method for building fuzzy clusters of large data sets, using a smoothing numerical approach. The usual sum-of-squares criterion is relaxed so the search for good fuzzy partitions is made on a continuous space, rather than a discrete space as in classical methods [2]. The smoothing allows a conversion from a strongly non-differentiable problem into low dimensional differentiable subproblems of optimization without constraints, by using an infinitely differentiable function. For the implementation of the algorithm we used the statistical software *R* and the results obtained were compared to the traditional fuzzy C–means method, proposed by Bezdek [1].

Keywords: clustering, fuzzy sets, numerical smoothing.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
2. Hartigan, J.A.: Clustering Algorithms. Wiley, New York, NY (1975)

David Masís
Costa Rica Institute of Technology, Cartago, Costa Rica, e-mail: dmasis@itcr.ac.cr

Esteban Segura
CIMPA & School of Mathematics, University of Costa Rica, San José, Costa Rica
e-mail: esteban.seguraugalde@ucr.ac.cr

Javier Trejos
CIMPA & School of Mathematics, University of Costa Rica, San José, Costa Rica
e-mail: javier.trejos@ucr.ac.cr

Adilson E. Xavier
Universidade Federal de Rio de Janeiro, Brazil, e-mail: adilson.xavier@gmail.com

Combining KDE and DBSCAN Clustering to Understand Road Traffic Accidents: the Case of Setúbal, Portugal

Pedro Nogueira, Marcelo Silva, Paulo Infante, Paulo Rebelo Manuel, Leonor Rego, Anabela Afonso, and Gonçalo Jacinto

Road traffic accidents (RTA) constitute a scourge that modern societies face, with an increasing death toll each passing year. Deep knowledge of the conditioning factors might help to mitigate this problem. Understanding the RTA location and what variables play a role are keys to foster road safety and outline prevention policies. The analysis of hotspots location based on RTA is the most common approach to understand the relations between neighboring accidents, looking for spatial significance. Herein, it is proposed that the comparison and analysis of hotspots with the clusters defined by DBSCAN algorithm is a valid tool to further clarify the spatial distribution of RTA. Data from the Portuguese district of Setúbal between the years 2016 and 2019 was used and the following datasets/subsets were defined: i) all accidents, ii) accidents with victims, and iii) accidents with fatalities and/or major injuries. The Kernel Density Estimation (KDE)was used with a quartic function to define the hotspots in QGIS.
The comparison of the hotspots with DBSCAN results allow us to conclude that: A) datasets i) and ii) have similar hotspot locations and there is no relation between hotspots and DBSCAN clusters, a single cluster comprises all the hotspots, being the other for small patches randomly distributed in the studied area; B) For dataset iii) the hotspots are not well defined, with one exception, whereas DBSCAN creates two clusters, separating urban areas with dense traffic, from more rural areas with traffic concentrated in high-speed roads; C) Moreover, for dataset iii) the remaining DBSCAN clusters define RTA in specific low traffic roads. These low traffic roads are, therefore, the targets that are prone to deepen the studies for understanding the location of RTA with fatalities or major injuries.

Pedro Nogueira and Marcelo Silva
ICT, Dep. de Geociências, Universidade de Évora,
e-mail: pmn@uevora.pt,marcelogs@uevora.pt

Paulo Infante, Anabela Afonso and Gonçalo Jacinto
CIMA, Dep. de Matemática, Universidade de Évora,
e-mail: pinfante@uevora.pt,aafonso@uevora.pt,gjcj@uevora.pt

Paulo Rebelo Manuel
CIMA, Universidade de Évora, e-mail: pjsrm@uevora.pt

Leonor Rego
Universidade de Évora, e-mail: lrego@uevora.pt

Similarity Forest for Time Series Classification

Tomasz Górecki, Maciej Łuczak, and Paweł Piasecki

The idea of similarity forest comes from Sathe and Aggarwal [1] and is derived from random forest. Random forests proved to be one of the most excellent methods, showing top performance across a vast array of domains, preserving simplicity, time efficiency, still being interpretable at the same time. However, its usage is limited to multidimensional data. Similarity forest does not require such representation — it is only needed to compute similarities between observations. Thus, it may be applied to data, for which multidimensional representation is not available. In this paper, we propose the implementation of similarity forest for time series classification. We compare the performance of similarity forest with 1NN classifier and random forest on the UCR benchmark database. We show that similarity forest with DTW, taking into account mean ranks, outperforms other classifiers. The comparison is enriched with statistical analysis.

Keywords: time series, time series classification, random forest, similarity forest

References

1. Sathe, S., Aggarwal, C. C.: Similarity Forests. Proc. of the 23rd ACM SIGKDD, 395–403 (2017)

Tomasz Górecki
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, Poznań, e-mail: tomasz.gorecki@amu.edu.pl

Maciej Łuczak
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, Poznań, e-mail: maciej.luczak@amu.edu.pl

Paweł Piasecki
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Uniwersytetu Poznańskiego 4, Poznań, e-mail: pawel.piasecki@amu.edu.pl

Uncovering Regions of Maximum Dissimilarity on Random Process Data

Miguel de Carvalho and Gabriel Martos Venturini

Everyday millions of data patterns flow around the world at unprecedented speed, thus leading to an explosion on the demand for modeling stochastic process data—such as time series, point processes, and functional data; each of these types of data plays a key role in machine learning, as can be seen, for instance, from the recent papers of [1], [2], and [3]. Hand in hand with this shock on demand arrived a pressing need for the development of data-intensive methods, techniques, and algorithms for *learning and comparing random processes*.

In this talk, I will propose a statistical method that learns about regions with a certain volume, where the marginal attributes of two processes are less similar. The proposed methods are devised in full generality for the setting where the data of interest are themselves stochastic processes, and thus the proposed method can be used for pointing out the regions of maximum dissimilarity with a certain volume, in the contexts of functional data, time series, and point processes. The parameter functions underlying both stochastic processes of interest are modeled via a basis representation, and Bayesian inference is conducted via an integrated nested Laplace approximation. The numerical studies validate the proposed methods, and we showcase their application with case studies on criminology, finance, and medicine.

Keywords: functional parameters, multi-objective optimization, pairs of random processes, Kolmogorov metric, set function optimization, Youden J statistic

References

1. José R Berrendero, Beatriz Bueno-Larraz, and Antonio Cuevas. On Mahalanobis distance in functional settings. J. Machine Learning Res., **21(9)cl**, 1–33 (2020).
2. Johann Faouzi and Hicham Janati. `pyts`: A python package for time series classification. J. Machine Learning Res., **21(46)**, 1–6 (2020).
3. Ganggang Xu, Ming Wang, Jiangze Bian, Hui Huang, Timothy R. Burch, Sandro C. Andrade, Jingfei Zhang, and Yongtao Guan. Semi-parametric learning of structured temporal point processes. J. Machine Learning Res., **21(192)**, 1–39 (2020).

Miguel de Carvalho
School of Mathematics, University of Edinburgh, UK, e-mail: `Miguel.deCarvalho@ed.ac.uk`

Gabriel Martos Venturini
Universidad Torcuato Di Tella, Buenos Aires, Argentina, e-mail: `gmartos@utdt.edu`

Franz Liszt's Transcendental Études: an Evolutionary Analysis by Machine Learning

Matteo Farnè

Musical data mining is a young discipline, that is gaining momentum in recent years (see [1]). In this paper, we apply the most relevant tools of Music Information Retrieval (MIR, see [2]) to Franz Liszt's Transcendental Études, with the aim to mark the evolution of his composition style and musical grammar. We consider the three versions of Transcendental Études published by the composer in 1826, 1837 and 1851. We perform a systematic evolutionary analysis of each Étude, and we compare different recordings of some Études.

For each trace, we estimate the amplitude spectrum, the envelope spectrum, and the spectrogram, in order to retrieve the musical content in terms of frequencies and intensity over time. Based on the estimated spectral features, we derive the chromagram, that is the redistribution of the spectrum over the twelve notes of the chromatic scale across all the registers. We also perform a segmentation based on the degree of novelty, intended as spectral dissimilarity, calculated frame-by-frame via the cosine distance. This process allows to discover and compare the macro-formal structure of the Études across the three published versions in terms of harmonic and melodic content.

Generally speaking, we learn that the first version represents a sketch of each Étude, the second version is a highly technical evolution of the first version, while the definitive version is characterized by the high degree of technical difficulty of the second version and the same formal clarity of the first version.

Keywords: musical data mining, spectral analysis, Franz Liszt

References

1. Cancino-Chacón, C., Carlos E., Grachten, M., Goebl, W., Widmer, G.: Computational models of expressive music performance: a comprehensive and critical review. Frontiers in Digital Humanities. **5**, 321–354 (2018)
2. Müller, M. Fundamentals of music processing: audio, analysis, algorithms, applications. Springer, 2015.
3. Lartillot, O., Toiviainen, P., Eerola, T.: A MATLAB toolbox for music information retrieval. In: Data analysis, machine learning and applications, pp. 261-268. Springer, (2008)

Matteo Farnè
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41,
e-mail: matteo.farne@unibo.it

Quantile-distribution Functions and Their Use for Classification

Edoardo Redivo, Cinzia Viroli, and Alessio Farcomeni

We develop a flexible parametric framework for the estimation of quantile functions. The method involves the specification of an analytical quantile distribution function for the data at hand [1]. We focus on quantile functions that are linear with respect to their parameters, such as the flattened generalized logistic distribution [2]: these can adapt to a wide range of distributional shapes and allow for the estimation to be carried out through a computationally efficient least-squares method based on the order statistics.
Inferential properties of this estimator, such as its asymptotic distribution, are derived, and these allow for the definition of a test of hypothesis for the equality of two distributions. The properties of the test are evaluated via a simulation study.
Our method of quantile function estimation is implemented as a density estimation method in the naïve Bayes classifier. This innovation is compared to standard approaches for the classifier in a simulation study, and is illustrated on a real data set coming from microRNA profiling in human Medulloblastoma. Moreover, the test of hypothesis is shown to be useful as a variable selection method.

Keywords: quantile function estimation, naïve bayes, variable selection

References

1. Gilchrist, W.: Statistical Modelling with Quantile Functions. Taylor & Francis, Andover (2000)
2. Chakrabarty, T.K., Sharma, D.: A generalization of the quantile-based flattened logistic distribution. Ann. Data. Sci. **8**, 603–627 (2021)

Edoardo Redivo · Cinzia Viroli
University of Bologna, via Belle Arti 41, 40126 Bologna, Italy,
e-mail: {edoardo.redivo,cinzia.viroli}@unibo.it

Alessio Farcomeni
University of Rome "Tor Vergata", Via Columbia 2, 00133 Rome, Italy
e-mail: alessio.farcomeni@uniroma2.it

Analysis of Gini Splitting Criterion and Comparison with Maximum Likelihood Rule

Amirah S. Alharthi and Charles C. Taylor

A commonly used criterion in decision trees is the Gini index. Considering random variables from two populations, with priors p_1 and p_2, the expected value of the Gini function is given by the asymptotic result: $(F_1(x) - F_2(x))^2/\{F(x)(1-F(x))\}$, where class i has distribution function $F_i(x)$ and $F(x) = p_1F_1(x) + p_2F_2(x)$ [1]. In this population setting, x would be chosen to maximize this. This result is obtained by taking the conditional expectation of the weighted Gini expression: $\sum_i N_{Li}^2/N_L + N_{Ri}^2/N_R$, in which the random variables N_{L1} denote the number in class 1 to the left of a split, etc. In contrast, the maximum likelihood (ML) classifier allocates according to $\arg\max(p_1f_1(x), p_2f_2(x))$, where $f_i(x)$ is the density of the ith population.

We consider the case of two normal populations, where (without loss of generality) $f_1(x)$ is the standard normal distribution, and $f_2(x)$ is normal with mean $\mu > 0$ and variance σ^2, to find cases in which the two splitting rules are the same, or differ. When $p_1 = p_2 = 1/2$ and $\sigma = 1$ both rules will split at $x = \mu/2$.

When $\sigma = 1$, then ML gives a split at $\mu/2 + \mu^{-1}\log(p_1/p_2)$, whereas an approximate solution for the Gini split, obtained by taking the derivative of the log of the above expected value, then taking a Taylor series expansion around $x = \mu/2$ and equating to zero, is: $\mu/2 + 2P^2\sqrt{2\pi}(p_2 - p_1)/\{4P\exp(-\mu^2/8) - \mu\sqrt{2\pi}\}$, where $P = 2\Phi(\mu/2) - 1$ and $\Phi(\cdot)$ is the CDF of the standard normal distribution. When $p_1 \neq p_2$, differences may be large, particularly as μ gets closer to 0. In this case when $\sigma = 1$, the Gini split is always in the interval $(0, \mu)$. When $p_1 \neq p_2$ we have not been able to obtain an approximation to the Gini solution for general σ. However, in some examples, it can be seen that the MLE split and Gini split are generally closer together and there are cases in which neither split is in the interval $(0, \mu)$.

Keywords: classification, Gini index, maximum likelihood

References

1. Alharthi, A.S.: Weighted Classification Tree-based Ensemble Methods. PhD thesis, University of Leeds, U.K. (2020)

Amirah · Charles C. Taylor
University of Leeds, Leeds LS2 9JT, UK, e-mail: c.c.taylor@leeds.ac.uk

Envelope-based Support Vector Machine Classifier

Alya Alzahrani and Andreas Artemiou

The envelope method is a relatively new and efficient dimension reduction technique that was introduced in the regression framework by Cook 2010 [2]. In this work, we extended this method to classification and developed a new projection-based approach based on a Support Vector Machine (SVM) classifier. Our proposed classifier is obtained by combining the envelope method and SVM to achieve a better and more efficient classification. Using the idea of the envelope to extract a lower-dimensional subspace projected the data on has advanced the classification performance.

Keywords: classification, dimension reduction, support vector machine, envelope methods.

References

1. Cook, R. Dennis, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica. 927-960 (2010)

Alya Alzahrani
School of Mathematics Cardiff University , Senghennydd Rd, Cardiff CF24 4AG,
e-mail: alzahranias4@cardiff.ac.uk

Andreas Artemiou
School of Mathematics Cardiff University, Senghennydd Rd, Cardiff CF24 4AG,
e-mail: artemioua@cardiff.ac.uk

A Moment-free Measure of Multivariate Skewness

Andrzej Sokolowski and Malgorzata Markowska

Multivariate skewness measures proposed in the literature are usually generalizations of univariate ones, and are based on the third central moment or Pearson's relation between mode, mean and standard deviation. Popular measure was proposed by Mardia [2], and good review of other propositions can be found in [1]. The aim of the paper is to propose a new measure of sample multivariate skewness which uses the idea of a "mirror observation". The measure is calculated on standardized data. The "mirror observation" is an artificial point lying on the line from the given data point through the coordinate origin, within the same distance to the origin as the actual point, and located on the other side of the origin. Then the distance from the real data point closest to the mirror one is used in the construction of the measure, which is finally the average over all data points. The problem of assymetry sign is discussed as well as the distribution of the measure under multivariate normal distribution. The empirical example compares the multivariate skewness of different socio-economic spheres in European Union countries distribution.

Keywords: multivariate skewness, nearest neighbor, european union countries

References

1. Balakrishnan N. and Scarpa B.: Multivariate measures of skewness for the skew-normal distribution. J. Multivariate Anal. 104, 73-87 (2012).
2. Mardia K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika 36, 519-530 (1970).

Andrzej Sokolowski
Cracow University of Economics, Rakowicka 27, 31-510 Kraków, Poland
e-mail: andrzej.sokolowski@uek.krakow.pl

Malgorzata Markowska
Wroclaw University of Economics and Business, Komandorska 118/120, Wroclaw, Poland
e-mail: malgorzata.markowska@ue.wroc.pl

The Weighted RV Coefficient: Exact Moments by Invariant Orthogonal Integration

François Bavaud

Weighted configurations $(\mathbf{f}, \mathbf{D})$, describing the squared Euclidean dissimilarities $\mathbf{D}$ between n objects endowed with a normalized vector of weights $\mathbf{f}$, are pervasive in Data Analysis. Weighted classical MDS, returning the configuration coordinates maximizing the low-dimensional proportion of inertia, obtains as a straightforward generalization of the well-known Torgerson-Gower procedure. It is based upon the spectral decomposition of the *matrix of weighted scalar products* or *kernel* $\mathbf{K}$.

Comparing two weighted configurations $(\mathbf{f}, \mathbf{D}_X)$ and $(\mathbf{f}, \mathbf{D}_Y)$ with identical weights $\mathbf{f}$ can be performed by computing the coefficient $\mathtt{CV}_{XY} = \mathrm{trace}(\mathbf{K}_X\mathbf{K}_Y)$, or its normalized version $\mathtt{RV}_{XY} = \mathtt{CV}_{XY}/\sqrt{\mathtt{CV}_{XX}\mathtt{CV}_{YY}} \in [0, 1]$, which constitutes the weighted extension of the RV similarity coefficient [1].

In the literature, there seems to be no complete agreement for the expressions of the null expectation of the RV first moments, permitting to asses the significance of the association between two configurations. We propose a new procedure consisting to integrate out products of orthogonal matrices occurring in the spectral decompositions of $\mathbf{K}_X$ and $\mathbf{K}_Y$, yielding exact expressions for the three first moments of the weighted RV coefficient; they depend on n and on the *spectral moments* of the eigenvalues of $\mathbf{K}_X$ and $\mathbf{K}_Y$ (their scree graphs), but not on $\mathbf{f}$.

Besides its relevance for applications (the scope of data analytic problems able to be expressed by various kernels, including conditional kernels, seems inexhaustible), the present approach sheds new light on some formal issues of interest, such as:

- Under the null distribution, the skewness of the RV coefficient is here proportional to the product of both spectral skewness, thus implying a positive RV skewness for most "natural" configurations, as often noticed in the literature [2].
- The traditional Moran test of spatial auto-correlation fits info the present framework, and its application can be generalized to multivariate features (and weighted regions) by the introduction of an exact variance-deflating correction.

Keywords: weighted Rv coefficient, permutation test, orthogonal Haar integration

References

1. Robert, P. and Escoufier, Y.: A unifying tool for linear multivariate statistical methods: the RV-coefficient. Journal of the Royal Statistical Society: Series C **25**(3), 257–265 (1976)
2. Josse, J., Pagès, J., and Husson, F.: Testing the significance of the RV coefficient. Computational Statistics & Data Analysis **53**(1), 82–91 (2008)

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

Testing Equality of Multivariate Coefficients of Variation

Marc Ditzhaus and Łukasz Smaga

The univariate coefficient of variation is well known unit-free variability measure, which is often applicable. There are a few its multivariate extensions, and no one is assumed to be a default [2]. Moreover, only for one of the multivariate coefficients of variation, statistical tests are known for verifying their equality in several groups [1, 3]. In this paper, we would like to fill this gap. We prove that the asymptotic distribution of the estimators of multivariate coefficients of variation is normal with appropriate variances. Using this result, we construct a Wald-type test statistics, whose distributions are approximated by the permutation method. The properties of the obtained tests are investigated in simulation studies. We consider the control of type I error level and power.

Keywords: multivariate coefficient of variation, permutation method, statistical test

References

1. Aerts, S., Haesbroeck, G.: Robust asymptotic tests for the equality of multivariate coefficients of variation. Test **26**, 163–187 (2017)
2. Albert, A., Zhang, L.: A novel definition of the multivariate coefficient of variation. Biom. J. **52**, 667–675 (2010)
3. Ditzhaus, M., Smaga, Ł.: Permutation test for the multivariate coefficient of variation in factorial designs. J. Multivariate Anal. **187**, 104848 (2022)

Marc Ditzhaus
Institute for Mathematics, Otto-von-Guericke University Magdeburg, Magdeburg, Germany,
e-mail: marc.ditzhaus@ovgu.de

Łukasz Smaga
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland,
e-mail: ls@amu.edu.pl

A New Regression Model for the Analysis of Microbiome Data

Roberto Ascari and Sonia Migliorati

Human microbiome data are becoming extremely common in biomedical research due to the relevant connections with different types of diseases. A widespread discrete distribution to analyze this kind of data is the Dirichlet-multinomial. Despite its popularity, this distribution often fails in modeling microbiome data due to the strict parameterization imposed on its covariance matrix.
The aim of this work is to propose a new distribution for analyzing microbiome data and to define a regression model based on it. The new distribution can be expressed as a structured finite mixture model with Dirichlet-multinomial components. We illustrate how this mixture structure can improve a microbiome data analysis to cluster patients into "enterotypes", which are a classification based on the bacteriological composition of gut microbiota. The comparison between the two models is performed through an application to a real gut microbiome dataset.

Keywords: count data, bayesian inference, mixture model, multivariate regression

Roberto Ascari
Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Milan, Italy; e-mail: roberto.ascari@unimib.it

Sonia Migliorati
Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Milan, Italy; e-mail: sonia.migliorati@unimib.it

The Death Process in Italy Before and During the Covid-19 Pandemic: a Functional Compositional Approach

Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli, and Piercesare Secchi

In this talk, based on [1], we propose a spatio-temporal analysis of daily death counts in Italy, collected by ISTAT (Italian Statistical Institute), in Italian provinces and municipalities. While in [1] the focus was on the elderly class (70+ years old), we here focus on the middle class (50-69 years old), carrying out analogous analyses and comparative observations. We analyse historical provincial data starting from 2011 up to 2020, year in which the impacts of the Covid-19 pandemic on the overall death process are assessed and analysed. The cornerstone of our analysis pipeline is a novel functional compositional representation for the death counts during each calendar year: specifically, we work with mortality densities over the calendar year, embedding them in the Bayes space B^2 of probability density functions. This Hilbert space embedding allows for the formulation of functional linear models, which are used to split each yearly realization of the mortality density process in a predictable and an unpredictable component, based on the mortality in previous years. The unpredictable components of the mortality density are then spatially analysed in the framework of Object Oriented Spatial Statistics. Via spatial downscaling of the results obtained at the provincial level, we obtain smooth predictions at the fine scale of Italian municipalities; this also enable us to perform anomaly detection, identifying municipalities which behave unusually with respect to the surroundings

Keywords: covid-19, o2s2, functional data analysis, spatial downscaling

References

1. Scimone, R., Menafoglio, A.,Sangalli, L.M., Secchi,P. (2021): A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. Spatial Statistics, doi: https://doi.org/10.1016/j.spasta.2021.100541, preprint available at `https://mox.polimi.it/reports-and-theses/publication-results/?id=952`

Riccardo Scimone · Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy, Center for Analysis, Decision and Society, Human Technopole
e-mail: `riccardo.scimone@polimi.it;piercesare.secchi@polimi.it`

Alessandra Menafoglio · Laura M. Sangalli
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
e-mail: `alessandra.menafoglio@polimi.it;laura.sangalli@polimi.it`

Sampling Design for Uncovering Natural Laws in Compositional Data

Lan Liang, Glòria Mateu-Figueras, and Jan Graffelman

Natural law refers to a stable mathematical relationship being constructed by some parts of the composition obtained from nature. One of the common forms of natural laws in the composition is the constant logcontrast relationship between parts, usually indicating a compositional system at an equilibrium state. A statistical approach to detecting this form of law includes two steps: 1) collecting appropriate data for the analysis, and 2) choosing proper statistical methods to analyze the data. In the aspect of statistical methodology, ample studies have shown that the logratio principal component analysis (LR-PCA) is an effective tool for detecting the constant logcontrast patterns [1, 2]. However, though the sampling design is also an important aspect of statistical design, limited literature discusses its effect on drawing accurate conclusions about the laws in composition. Therefore, this study aims to investigate the limitations of different sampling methods on LR-PCA procedure and develop more suitable sampling strategies for law detection.

Through a simulation study of genotype frequencies in Hardy Weinberg Equilibrium (HWE), we found that when generating a 3D compositional data set containing HWE law, if one variable is assigned an exact value and thus has a small variance, the ratios between pair-wise variables are close to constants. In this context, the LR-PCA may report the proportionality between two parts instead of the law of interest. To avoid this problem, high heterogeneity in compositional samples is required.

Keywords: compositional data, natural law, constant logcontrast, logratio principal component analysis (LR-PCA), sampling

References

1. Aitchison, J.: Logratios and natural laws in compositional data analysis. Mathematical Geology. **31 (5)**, 563–580 (1999)
2. Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: Modeling and Analysis of Compositional Data. John Wiley & Sons, Chichester (2015)

Lan Liang
Technical University of Catalonia, Barcelona, Spain, e-mail: lan.liang@upc.edu

Glòria Mateu-Figueras
University of Girona, Girona, Spain, e-mail: gloria.mateu@udg.edu

Jan Graffelman
Technical University of Catalonia, Barcelona, Spain, e-mail: jan.graffelman@upc.edu
University of Washington, Seattle, USA

Penalized Model-based Functional Clustering: a Regularization Approach via Shrinkage Methods

Nicola Pronello, Rosaria Ignaccolo, Luigi Ippoliti, and Sara Fontanella

With the advance of modern technology, and with data being recorded continuously, functional data analysis has gained a lot of popularity in recent years. Working in a mixture model-based framework, we develop a flexible functional clustering technique achieving dimensionality reduction schemes through a L_1 penalization. The proposed procedure results in an integrated modelling approach where shrinkage techniques are applied to enable sparse solutions in both the means and the covariance matrices of the mixture components, while preserving the underlying clustering structure. This leads to an entirely data-driven methodology suitable for simultaneous dimensionality reduction and clustering. Preliminary experimental results, both from simulation and real data, show that the proposed methodology is worth considering within the framework of functional clustering.

Keywords: functional data analysis, l_1 penalty, silhouette width, graphical lasso, mixture model.

Nicola Pronello
Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Chieti, Italy, e-mail: nicola.pronello@unich.it

Rosaria Ignaccolo
Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Torino, Italy, e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
Department of Economics, University of Chieti-Pescara, Pescara, Italy
e-mail: luigi.ippoliti@unich.it

Sara Fontanella
National Heart and Lung Institute, Imperial College London, London, United Kingdom
e-mail: s.fontanella@imperial.ac.uk

Clustering in FDA Mixing the Epigraph and the Hypograph Indexes with Machine Learning Algorithms

Belén Pulido, Alba M. Franco-Pereira, and Rosa E. Lillo

Clustering is considered as one of the most used techniques in Data Science. Clustering functional data is a challenging problem since it involves working in an infinite dimensional space. In this work this problem is addressed by applying the epigraph and the hypograph indexes to a functional dataset and thereby, converting it from a functional data problem into a multivariate problem. See [1]. Once the multivariate dataset is obtained, the techniques that have been fully studied in the literature for clustering multivariate data can be applied, including both procedures typical of the area of Statistics and those from the machine learning field. This methodology is applied to both simulated and real datasets, and it is also compared to two clustering techniques originally designed for functional data ([2] and [3]).

Keywords: epigraph, hypograph, clustering, functional data, machine learning algorithms

References

1. Pulido, B., Franco-Pereira, A.M., Lillo, R.E.: Functional clustering via multivariate clustering. arXiv:2108.00217 [stat.ME] (2021)
2. Martino, A., Ghiglietti, A., Ieva, F., Paganoni, A. M.: A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data. Statistical Methods and Applications, **28(2)**, 301-322 (2019)
3. Zambom, A. Z., Collazos, J. A., Dias, R.: Functional data clustering via hypothesis testing k-means. Computational Statistics, **34(2)**, 527-549 (2019)

Belén Pulido
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain,
e-mail: belen.pulido@uc3m.es

Alba M. Franco-Pereira
Department of Statistics and O.R., Universidad Complutense de Madrid, Madrid, Spain,
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain,
e-mail: albfranc@ucm.es

Rosa E. Lillo
Department of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid, Spain,
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain,
e-mail: rosaelvira.lillo@uc3m.es

Localization Processes for Functional Data Classification

Antonio Elías, Raúl Jiménez, and Joseph E. Yukich

We propose an alternative to k-nearest neighbors for functional data whereby the approximating neighboring curves are piecewise functions built from a functional sample. Using a locally defined distance function that satisfies stabilization criteria, we establish pointwise and global approximation results in function spaces when the number of data curves is large enough. We exploit this feature to develop the asymptotic theory when a finite number of curves is observed at time-points given by an i.i.d. sample whose cardinality increases up to infinity. We use these results to study the problem of functional classification and outlier detection. For such problems our methods are competitive with and sometimes superior to benchmark predictions in the field.

Keywords: functional data classification, nearest neighbors, outlier detection

Antonio Elías
Department of Applied Mathematics, Universidad de Málaga, Spain,
e-mail: aelias@uma.es

Raúl Jiménez
Department of Statistics, Universidad Carlos III, Madrid, Spain,
e-mail: rauljose.jimenez@uc3m.es

J. E. Yukich
Department of Mathematics, Lehigh University, USA,
e-mail: joseph.yukich@lehigh.edu

A New Functional Data Clustering Technique Based on Spectral Clustering and Downsampling

Maryam Al Alawi, Surajit Ray, and Mayetri Gupta

We present a new framework for clustering functional data along with a new paradigm for performing model selection based on downsampling. Our clustering framework is a generalisatiion of the spectral clustering approach and is flexible enough to exploit higher order features of curves, including derivatives. Extensive comparative studies with existing methods show a clear advantage of our approach over existing functional data analysis clustering approaches. Additionally, we present a new paradigm for model selection, by introducing the technique of downsampling, which allows us to create lower resolution replicates of the observed curves. These replicates can then be used to provide insight into the tuning parameters for the specific clustering techniques. The usefulness of the proposed methods is illustrated through simulations and applications to real-life datasets.

Keywords: clustering, clustering stability, functional data analysis, model selection, functional data clustering

References

1. Ramsay, JO and Silverman, BW.: Functional Data Analysis. Springer, c1997. New York (2005)
2. Ng, Andrew Y and Jordan, Michael I and Weiss, Yair.: On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems. 849–856 (2002)
3. Von Luxburg, Ulrike.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
4. Al Alawi, M., Ray, S. and Gupta, M. A New Framework for Distance-based Functional Clustering. In: 34th International Workshop on Statistical Modelling, Guimarães, Portugal, 07-12 Jul 2019, (2019)

Maryam Al Alawi
Sultan Qaboos University, Oman e-mail: malalawi@squ.edu.om

Surajit Ray
University of Glasgow, United Kingdom, e-mail: surajit.ray@glasgow.ac.uk

Mayetri Gupta
University of Glasgow, United Kingdom, e-mail: mayetri.gupta@glasgow.ac.uk

Parsimonious Mixtures of Seemingly Unrelated Contaminated Normal Regression Models

Gabriele Perrone and Gabriele Soffritti

In recent years, the research into multivariate linear regression based on finite mixture models has been intense. With such an approach, it is possible to perform regression analysis for a multivariate response by taking account of the possible presence of several unknown homogeneous groups, each of which is characterised by a different linear regression model. For a continuous multivariate response, mixtures of normal regression models are generally employed. However, in real data, mildly atypical observations can negatively affect the estimation of the regression parameters under a normal distribution in each mixture component. Robust methods insensitive to the presence of such observations have been recently introduced [1]. Furthermore, in some fields of research, a multivariate regression model with a different vector of covariates for each response should be specified, based on some prior information to be conveyed in the analysis. This approach has been recently embedded into the framework of Gaussian mixture models [2]. To take account of all these aspects, mixtures of seemingly unrelated contaminated normal regression models has been defined [3]. A further extension is presented here so as to ensure parsimony, which is obtained by imposing constraints on the component-covariance matrices. The resulting parsimonious mixtures of seemingly unrelated contaminated regression models are described together with an illustration of their practical usefulness.

Keywords: contaminated normal distribution, ecm algorithm, mixture of regression models, model-based cluster analysis, seemingly unrelated regression.

References

1. Mazza, A., Punzo, A.: Mixtures of multivariate contaminated normal regression models. Stat. Pap. **169**, 787–822 (2020)
2. Galimberti, G., Soffritti, G.: Seemingly unrelated clusterwise linear regression. Adv. Data Anal. Classif. **14**, 235–260 (2020)
3. Perrone, G., Soffritti, G.: Seemingly unrelated clusterwise linear regression for contaminated data. Under review (2021)

Gabriele Perrone
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy, e-mail: gabriele.perrone4@unibo.it

Gabriele Soffritti
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy, e-mail: gabriele.soffritti@unibo.it

Monitoring Hyperparameter Choice for Robust Cluster Weighted Model

Andrea Cappozzo, Luis A. García-Escudero, Francesca Greselin, and Agustín Mayo-Iscar

The estimation of the Cluster Weighted Model is particularly attractive for providing explicit modeling of the explanatory variables, in a mixture of regression. The Robust version of the model requires the specification of a set of crucial parameters, like the proportion of trimmed units α, the thresholds to be adopted for the constrained estimation of groups scatter and for regression errors, beyond the number of components of the Mixture. To assist the choice of such hyper-parameters, a monitoring methodology could be of great help. The purpose is to provide a set of graphical tools to guide the final user in making an informed judgment, considering a landscape of plausible choices. The final output offers a set of optimal solutions, featured by the interval of hyper-parameters values in which their optimality holds, their stability and validity. An assessment of the role and extent of the outlying observations has been provided, introducing three new silhouette plots. The purpose is to understand the possible effects of the contaminated observations, with respect to the clustering of the covariate X, and the local regression lines Y, following the nature of the Cluster Weighted model.

Keywords: cluster-weighted modeling, outliers, trimmed bic, eigenvalue constraint, monitoring, model-based clustering, robust estimation

References

1. Cappozzo, A., García Escudero, L.A., Greselin, F., and Mayo-Iscar, A.: Parameter Choice, Stability and Validity for Robust Cluster Weighted Modeling. Stats **4** (3), 602–615 (2021)
2. Riani, M., Atkinson, A. C., Cerioli, A., and Corbellini, A.: Efficient robust methods via monitoring for clustering and multivariate data analysis. Pattern Recognit. **88**, 246?260, (2019)
3. Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

Andrea Cappozzo
Department of Mathematics, Politecnico di Milano, e-mail: andrea.cappozzo@polimi.it

Luis A. García Escudero
Departamento de Estadística e Investigación Operativa, Universidad de Valladolid,
e-mail: lagarcia@uva.es

Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
e-mail: francesca.greselin@unimib.it

Agustín Mayo-Iscar
Departamento de Estadística e Investigación Operativa, Universidad de Valladolid,
e-mail: agustin.mayo.iscar@uva.es

Latent Block Regression Model

Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif

When dealing with high dimensional sparse data, such as in recommender systems, co-clustering turns out to be more beneficial than one-sided clustering, even if one is interested in clustering along one dimension only. Thereby, co-clusterwise is a natural extension of clusterwise. Unfortunately, all of the existing approaches do not consider covariates on both dimensions of a data matrix. In this paper, we propose a *Latent Block Regression Model* (LBRM) overcoming this limit. For inference, we propose an algorithm performing simultaneously co-clustering and regression where a linear regression model characterizes each block. Placing the estimate of the model parameters under the maximum likelihood approach, we derive a Variational Expectation-Maximization (`VEM`) algorithm for estimating the model's parameters. The finality of the proposed `VEM-LBRM` is illustrated through simulated datasets.

Keywords: co-clustering, clusterwise, tensor, data mining

Rafika Boutalbi
Institute for Parallel and Distributed Systems, Analytic Computing, University of Stuttgart, Germany, e-mail: `rafika.boutalbi@ipvs.uni-stuttgart.de`

Lazhar Labiod · Mohamed Nadif
Université de Paris, CNRS, Centre Borelli UMR 9010, Paris, France
e-mail: `lazhar.labiod@u-paris.fr;mohamed.nadif@u-paris.fr`

Towards a Bi-stochastic Matrix Approximation of *k*-means and Some Variants

Lazhar Labiod and Mohamed Nadif

The k-means algorithm and some k-means variants have been shown to be useful and effective to tackle the clustering problem. In this paper we embed k-means variants in a bi-stochastic matrix approximation (BMA) framework. Then we derive from the k-means objective function a new formulation of the criterion. In particular, we show that some k-means variants are equivalent to algebraic problem of bi-stochastic matrix approximation under some suitable constraints. For optimizing the derived objective function, we develop two algorithms; the first one consists in learning a bi-stochastic similarity matrix while the second seeks for the optimal partition which is the equilibrium state of a Markov chain process. Numerical experiments on real data-sets demonstrate the interest of our approach..

Keywords: k-means, reduced k-means, factorial k-means, bi-stochastic matrix

Lazhar Labiod
Université de Paris, CNRS, Centre Borelli UMR 9010, e-mail: lazhar.labiod@u-paris.fr

Mohamed Nadif
Université de Paris, CNRS, Centre Borelli UMR 9010, e-mail: mohamed.nadif@u-paris.fr

Clustering Brain Connectomes Through a Density-peak Approach

Riccardo Giubilei

The density-peak (DP) algorithm is a mode-based clustering method that identifies cluster centers as data points being surrounded by neighbors with lower density and far away from points with higher density. Since its introduction in 2014, DP has reaped considerable success for its favorable properties. A striking advantage is that it does not require data to be embedded in vector spaces, potentially enabling applications to arbitrary data types. In this work, we propose improvements to overcome two main limitations of the original DP approach, i.e., the unstable density estimation and the absence of an automatic procedure for selecting cluster centers. Then, we apply the resulting method to the increasingly important task of graph clustering, here intended as gathering together similar graphs. Potential implications include grouping similar brain networks for ability assessment or disease prevention, as well as clustering different snapshots of the same network evolving over time to identify similar patterns or abrupt changes. We test our method in an empirical analysis whose goal is clustering brain connectomes to distinguish between patients affected by schizophrenia and healthy controls. Results show that, in the specific analysis, our method outperforms many existing competitors for graph clustering.

Keywords: nonparametric statistics, mode-based clustering, networks, graph clustering, kernel density estimation.

Riccardo Giubilei
Luiss Guido Carli, Rome, Italy, e-mail: rgiubilei@luiss.it

New Metrics for Classifying Phylogenetic Trees Using *k*-means and the Symmetric Difference Metric

Nadia Tahiri and Aleksandr Koshkarov

The k-means method can be adapted to any type of metric space and is sometimes linked to the median procedures. This is the case for symmetric difference metric (or Robinson and Foulds [1]) distance in phylogeny, where it can lead to median trees as well as to Euclidean Embedding. We show how a specific version of the popular k-means clustering algorithm, based on interesting properties of the Robinson and Foulds topological distance, can be used to partition a given set of trees into one (when the data is homogeneous) or several (when the data is heterogeneous) cluster(s) of trees. We have adapted the popular cluster validity indices of Silhouette, and *Gap* to tree clustering with k-means based on a previous work by Tahiri *et al.* [2]. In this article, we will show results of this new approach on a real dataset (aminoacyl-tRNA synthetases) of Woese *et al.* [3]. The new version of phylogenetic tree clustering makes the new method well suited for the analysis of large genomic datasets.

Keywords: clustering, symmetric difference metrics, k-means, phylogenetic trees, cluster validity indices

References

1. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences*. **53**, 131-147 (1981)
2. Tahiri, N., Willems, M. & Makarenkov, V. A new fast method for inferring multiple consensus trees using k-medoids. *BMC Evolutionary Biology*. **18**, 1-12 (2018)
3. Woese, C., Olsen, G., Ibba, M. & Soll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology And Molecular Biology Reviews*. **64**, 202-236 (2000)

Nadia Tahiri
Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada,
e-mail: Nadia.Tahiri@USherbrooke.ca

Aleksandr Koshkarov
Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K2R1, Canada;
Center of Artificial Intelligence, Astrakhan State University, Astrakhan, 414056, Russia
e-mail: Aleksandr.Koshkarov@USherbrooke.ca

Alternating Optimization Framework for Sparse Simultaneous Component Analysis Based on Data Integration

Rosember Guerra-Urzola, Juan C. Vera, Katrijn Van Deun, and Klaas Sijtsma

Given multiple data blocks from different sources sharing the same observations (such as psychological questionnaires or genetic risk scores), Simultaneous Component Analysis (SCA) aims to find a few linear combinations of the variables that explain as much as possible the variability in the joined data set. However, rooting the analysis on all variables makes interpretability difficult, especially in high-dimensional settings. Therefore, looking for a sparse structure is natural; it identifies the common and distinctive source of variation across all data blocks. Solving the sparse SCA problem is intractable, given its combinatorial nature. Here, the nonconvex SCA problem is formulated as different convex maximization problems over the sphere, inducing sparsity via cardinality constraint and lasso penalties. To solve these models, optimization algorithms based on the alternating directions methods are proposed; these algorithms find high-quality feasible solutions for large dimensions. Extensive experiments, including a real-world data set, are used to assess the solution quality, computational time, and scalability of the methods.

Keywords: alternating optimization, dimension reduction, simultaneous component analysis

Rosember Guerra-Urzola
Tilburg University, Warandelaan 2, 5037 AB Tilburg
e-mail: R.I.GuerraUrzola@tilburguniversity.edu

Juan C. Vera
Tilburg University, Warandelaan 2, 5037 AB Tilburg
e-mail: J.C.VeraLizcano@tilburguniversity.edu

Katrijn Van Deun
Tilburg University, Warandelaan 2, 5037 AB Tilburg,
e-mail: K.VanDeun@tilburguniversity.edu

Klaas Sijtsma
Tilburg University, Warandelaan 2, 5037 AB Tilburg,
e-mail: K.Sijtsma@tilburguniversity.edu

Joint Sparse Principal Component Analysis

Katrijn Van Deun

Comparing multivariate relations between different groups forms the core of many studies in the empirical sciences. Latent variable approaches such as principal component and factor analysis are most useful (and used) to explore such multivariate relations. The loadings are key to the interpretation of these latent variable models as they express the strength of association of the observed variables with the latent variables. Preferably variables load on one or a few components/factors only and have zero loadings elsewhere as this eases interpretation. In addition, when comparing multiple groups, also a clear distinction between those variables that function in the same way over groups and those that do not is needed: Loadings should be exactly equal between those groups where the variables function in the same way and unequal elsewhere. In this paper we propose a multigroup latent variable model, called joint sparse principal component analysis, that has these properties. Sparsity is imposed using cardinality constraints while equal loadings are obtained as the result of a fusion penalty. We efficiently solve the estimation problem by use of an alternating optimization procedure that includes the alternating direction method of multipliers (ADMM) as one of the steps. Tuning of the cardinality and fusion penalty is based on the index of sparseness. We illustrate with an example on the co-occurrence of experienced symptoms by cancer survivors belonging to different tumor types.

Keywords: multigroup data, regularized PCA, ADMM

Katrijn Van Deun
Tilburg University, Warandelaan 2, e-mail: k.vandeun@uvt.nl

Joint Sparse Principal Component Analysis: a Simulation Study

Tra Le and Katrijn Van Deun

Measurement invariance is of great importance in the social and behavioral sciences, as it allows for generalization of latent constructs across different groups, typically by investigating the equality of factor structures. Traditionally, in settings where the loadings for the different groups are not known beforehand, exploratory factor analysis is commonly used. However, it has several drawbacks, including that most methods cannot handle data with fewer observations than variables and other problems (subjective thresholds for loading differences, unrealistic assumptions, instability with small sample size, large amount of computational sources needed, etc.). To overcome these drawbacks, joint sparse principal component analysis (joint SPCA) has been proposed, which adopts a regularized and cardinality constrained least-square approach. The aim of this paper is to compare it with the best available EFA method, namely multigroup factor rotation (MGFR) [1] A simulation study was carried out to evaluate the performance of joint SPCA in comparison with the MGFR technique, on three types of performance measures: recovery rate of the zero/non-zero pattern in the loadings, Tucker's congruence, and computation time. Following the setup by [1], we varied the number of groups, group sizes, number of components, type and size of loading differences, and the number of loading differences. Based on the first two measures, joint SPCA performed slightly less well than MGFR which reported a goodness-of-loading-recovery statistic for optimally rotated loadings of .99. Averaged across 6000 simulated datasets, joint SPCA had a recovery rate and Tucker congurence of .96 (SD = .06) and .98 (SD = .02), respectively. The CPU time increased as the conditions got more complex. Averaged across 50 replications for each condition, the shortest time was 5.3s (2 groups) and the longest time was 21.3s (4 groups) on an i5 processor with 8GB RAM.

Keywords: multigroup, pca, measurement invariance

References

1. De Roover, K., Vermunt, J.: On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. Struct. Equ. Model. **26**, 905–923 (2019)

Tra Le
Tilburg University, Tilburg, The Netherlands, e-mail: t.t.le_1@tilburguniversity.edu

Katrijn Van Deun
Tilburg University, Tilburg, The Netherlands, e-mail: k.vandeun@tilburguniversity.edu

Copula-based Non-metric Unfolding on Augmented Data Matrix

Marta Nai Ruscone and Antonio D'Ambrosio

A multidimensional unfolding technique [1] that is not prone to degenerate solutions and is based on multidimensional scaling of a complete data matrix is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using Copulas-based association measures [2] among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). The proposed technique leads to acceptable recovery of given preference structures. Application on real datasets show that our procedure returns non-degenerate unfolding solutions.

Keywords: copulas, unfolding, multidimensional scaling

References

1. Cox, T.F. and Cox, M.A.A.: Multidimensional scaling. Chapman & hall/CRC (2000)
2. Nelsen, R.B.: An introduction to copulas. Springer, New York (2013)

Marta Nai Ruscone
University of Genoa, Genoa, e-mail: marta.nairuscone@unige.it

Antonio D'Ambrosio
University of Naples Federico II, Naples, e-mail: antdambr@unina.it

Emotion Classification Based on Single Electrode Brain Data: Applications for Assistive Technology

Duarte Rodrigues, Luis Paulo Reis, and Brígida Mónica Faria

This research case focused on the development of an emotion classification system aimed to be integrated in projects committed to improve assistive technologies. An experimental protocol was designed to acquire an electroencephalogram (EEG) signal that translated a certain emotional state. To trigger this stimulus, a set of clips were retrieved from an extensive database of pre-labeled videos[1]. Then, the signals were properly processed, in order to extract valuable features [2] and patterns to train the machine and deep learning models.There were suggested 3 hypotheses for classification: recognition of 6 core emotions; distinguishing between 2 different emotions and recognising if the individual was being directly stimulated or merely processing the emotion. Results showed that the first classification task was a challenging one, because of sample size limitation. Nevertheless, good results were achieved in the second and third case scenarios (70% and 97% accuracy scores, respectively) through the application of a recurrent neural network.

Keywords: emotions, brain-computer interface, eeg, machine/deep learning

References

1. Cowen, A., Keltner, D.: :Self-report captures 27 distinct categories of emotion bridged by continuous gradients In: Proceedings of the National Academy of Sciences of the United States of America (2017) doi: 10.1073/pnas.1702247114.
2. Jenke, R., Peer, A., Buss, M.:Feature Extraction and Selection for Emotion Recognition from EEG. In: IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 327-339, (2014) doi: 10.1109/TAFFC.2014.2339834

Duarte Rodrigues
Faculty of Engineering of University of Porto (FEUP), Rua Dr. Roberto Frias, s/n 4200-465 Porto Portugal, e-mail: up201705420@fe.up.pt

Luis Paulo Reis
Faculty of Engineering of University of Porto (FEUP) and Artificial Intelligence and Computer Science Laboratory (LIACC), Rua Dr. Roberto Frias, s/n 4200-465 Porto Portugal
e-mail: lpreis@fe.up.pt

Brígida Mónica Faria
School of Health, Polytechnic of Porto (ESS-P.PORTO) and Artificial Intelligence and Computer Science (LIACC), e-mail: monica.faria@ess.ipp.pt

On the Role of Data, Statistics and Decisions in a Pandemic

Ursula Garczarek, Beate Jahn, Sarah Friedrich, Joachim Behnke, Joachim Engel, Ralf Münnich, Markus Pauly, Adalbert Wilhelm, Olaf Wolkenhauer, Markus Zwick, Uwe Sieber, and Tim Friede

A pandemic poses particular challenges to public health decision-making because of the need to continuously adapt public measures to rapidly changing evidence and data availability. This presentation provides an overview of the process of decision making using data in a pandemic and gives recommendations for the different steps from a statistical perspective. A range of modelling techniques with different goals including mathematical, statistical and decision-analytic models applied in the COVID-19 context are briefly introduced. We discuss the importance of statistical literacy, and of effective dissemination and communication of findings.

One recommendation relates to the need and value of interdisciplinary cooperation. Cooperation is central in a pandemic to the society at large, but also specifically within the field of data science: we should act as a specialist group rather than as individuals, broadly positioned and media-sensitive. The presentation is based on a manuscript summarizing the discussion of an interdisciplinary group [2].

Presenting this topic at the IFCS, we aim to foster the understanding of the goals of these modelling approaches and the specific data requirements that are essential for data collection and transformation, the interpretation of results and for successful interdisciplinary collaborations among statisticians, epidemiologists, public health experts, social sciences, and ethicists, as well as health decision and communication scientists.

Keywords: pandemic, modelling, statistical literacy, decision making

References

1. Jahn, B., Friedrich, S., Behnke, J., Engel, J., Garczarek, U., Münnich, R., Pauly, M., Wilhelm, A., Wolkenhauer, O., Zwick, M., Siebert, U., Friede, T: On the role of data, statistics and decisions in a pandemic AStA Adv Stat Anal (forthcoming). https://doi.org/10.48550/arXiv.2108.04068

Ursula Garczarek
Cytel Inc, 675, Massachusetts Avenue, Cambridge, MA 02139 USA
e-mail: ursula.garczarek@cytel.com

A Deep Learning Analytics to Detect Dental Caries

Taerim Lee

Deep Learning Analytics uses predictive models that provide actionable information for a better prognosis of dental caries. It is a multidisciplinary approach based on dental caries data processing, AI technology-learning enhancement, dental caries data mining, and visualization. Three key components need further clarification to help them effectively apply deep learning in dental caries prognosis to explain the methods for conducting deep learning, the benefits of using deep learning, and the challenges of using learning analytics in dental caries. Discover significant socio-demographic factors and microbiologic factors to detect the prognosis of dental caries. Compare the efficiency with other prognosis models using support vector machine, linear discriminant, random forest, logistic regression by ROC curves using ICD-9 codes for dental caries, 365 boys and 340 girl cohort with dental caries and normal group together. All available data variables required to develop and test models were identified from a sociodemographic and microbiological records database. Data on 500 records among 705 was utilized for the development of the model and on 205 patients utilized to perform cross-validation analysis of the models. Socio-demographic data such as presenting signs & symptoms, presence of caries, microbiologic data, and corresponding diagnosis and outcomes were collected. Dental data was collected for each target group was utilized to retrospectively ascertain optimal preventive management for dental caries. Clinical presentations and corresponding treatment were utilized as training examples.

Keywords: deep learning analytics; support vector machine; random forest

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012)
2. Sutskever, I. Vinyals, O. & Le. Q. V. Sequence to sequence learning with neural networks. In Proc. Advances in Neural Information Processing Systems 27 3104–3112 (2014)

Taerim Lee
Department of Statistics & Data Science, Korea National Open University, Seoul, Republic of Korea, e-mail: trlee@knou.ac.kr

Identification of Shared Genetic Loci Between Psychiatric Disorders and Telomere Length and Evaluation of Their Role as Potential Drug Targets

Claudia Pisanu, Anna Meloni, and Alessio Squassina

Patients with psychiatric disorders such as bipolar disorder (BD), schizophrenia (SCZ) and major depression (MD) show features suggestive of accelerated cellular aging such as shorter telomere length (TL). However, contrasting results have also been reported. We leveraged large genome-wide association studies to investigate whether shared genetic factors might predispose patients to cellular aging or rather play a counteractive role. For BD (41,917 cases, 371,549 controls), SCZ (69,369 cases, 236,642 controls) and MD (170,756 cases, 329,443 controls) we used datasets from the Psychiatric Genomics Consortium, while for TL a meta-analysis including 78,592 individuals. We identified shared genetic loci with conjunctional false discovery rate (conjFDR) [1]. Heritability and bivariate local genetic correlation was investigated with LAVA, while target druggability with different tools, including DGIdb. We identified two loci shared between BD and TL: 1) lead single nucleotide polymorphism (SNP) rs113833990, conjFDR=0.03; 2) lead SNP rs12919664, conjFDR=0.002. The latter showed significant heritability for BD (h2=0.0007, p=5.8E-06) and TL (h2=0.0004, p=0.006) and significant local genetic correlation (rg=0.78, p=0.005). One locus shared between SCZ and TL (lead SNP rs143773357, conjFDR=0.03) showed significant heritability (SCZ: h2=0.002, p=1.1E-09; TL: h2=0.0004, p=0.044) and local genetic correlation (rg=0.78, p=0.007). For all loci, the lead SNP was associated with increased TL and predisposition to psychiatric disorders. Our results suggest that shorter TL in patients with SCZ or BD could be at least partly counteracted by genetic factors.

Keywords: genetic correlation, pleiotropy, genomics, psychiatry

References

1. Smeland, O.B., Frei, O., Shadrin, A. et al.: Discovery of shared genomic loci using the conditional false discovery rate approach. Hum. Genet. **139**, 85–94 (2020)

Claudia Pisanu
University of Cagliari, Cagliari, Italy, e-mail: claudia.pisanu@unica.it

Anna Meloni
University of Cagliari, Cagliari, Italy, e-mail: anna.meloni@unica.it

Alessio Squassina
University of Cagliari, Cagliari, Italy, e-mail: squassina@unica.it

Estimating Optimal Decision Trees for Treatment Assignment with $k > 2$ Treatment Alternatives: a Classification Problem with a Unit- and Class- dependent Misclassification Cost

Iven Van Mechelen and Aniek Sies

For many medical and psychological problems, multiple treatment alternatives are available. Given data from a randomized controlled trial, an important challenge is to estimate an optimal decision rule that specifies for each patient the most effective treatment alternative given his or her pattern of pretreatment characteristics. At this point, optimality refers to the most favorable expected (potential) outcome if the rule would be applied to the entire population of patients of interest. The estimation problem at hand can be shown to come down to a classification problem with a unit- and class-dependent misclassification cost, that is, a misclassification cost that may depend on both the object that is misclassified and the class to which it is erroneously assigned.
Classification trees constitute an insightful class of solutions for problems of decision rule estimation. Unfortunately, however, there is dearth of software tools for tree estimation that minimizes an object- and class-dependent misclassification cost, in particular for problems with $k > 2$ classes. In this talk, we explain how such an estimation can be achieved by means of a shrewd and novel type of application of a mainstream R-package for tree building, `rpart`, via a user-defined splitting function and a rectangular misclassification cost matrix. We illustrate with an application on the search for an optimal tree-based treatment regime in a randomized controlled trial on $k = 3$ different types of after-care for younger women with early-stage breast cancer. We finally argue that the proposed software solution may have relevance for various other classification problems with a unit- and class-dependent misclassification cost, such as credit card fraud detection and customer retention management.

Keywords: classification trees, unit- and class-dependent misclassification cost, optimal treatment regimes

Iven Van Mechelen
University of Leuven, Tiensestraat 102 - box 3713, 3000 Leuven, Belgium
e-mail: Iven.VanMechelen@kuleuven.be

Aniek Sies
University of Leuven, Tiensestraat 102 - box 3713, 3000 Leuven, Belgium
e-mail: Aniek.Sies@kuleuven.be

ExactTree: an R-package for Globally Optimal Decision Trees

Elise Dusseldorp, Juan Claramunt Gonzales, Jacqueline Meulman, Samil Uysal, and Bart Jan van Os

Decision trees, such as Classification and Regression Trees (CART) are grown using binary recursive partioning. The goal is to predict a categorical outcome (classification) or a continuous outcome (regression) as good as possible using binary splits on predictor variables. Because our goal is to preserve interpretability, we focus in this paper on single trees. The tree algorithm starts with all objects in the root node and subsequently searches for the predictor variable and split point that leads to the maximum decrease in impurity (e.g., residual sum of squares); then the root node is split into two child nodes. This process is repeated at each node until a full tree is grown. A downside of this recursive procedure is the risk of arriving at a local minimum. Therefore, several attempts have been made to grow globally optimal trees, among which evolutionary trees [1], based on a meta-heuristic algorithm, and the method ExactTree [2, 3], that optimizes the entire tree structure globally using dynamic programming. We performed a benchmark study comparing both methods on predictive accuracy and stability. Results on part of the data sets showed similar predictive accuracies, but higher stability for ExactTree. In our presentation, we show the final results and demonstrate the R-package ExactTree for you.

Keywords: decision trees, interpretable machine learning, classification, regression, global optimization

References

1. Grubinger, T., Zeileis, A., Pfeiffer, K-P: evtree: Evolutionary learning of globally optimal classification and regression trees in R. J. Stat. Softw. **61**, 1–29 (2014)
2. Os, B.J.: Dynamic Programming in Multivariate Data Analysis. Leiden University (2000)
3. Meulman, J.J., Dusseldorp, E., Os, B.J.: An exact dynamic programming algorithm for regression trees. In: Van der Heijden, M., Koren, B., Van der Mei, R.D., Van Vonderen, J.A.J. (eds.) Jan Karel Lenstra, the Traveling Science Man: Liber Amicorum, pp. 198–208. CWI, Amsterdam (2011).

Corresponding author: E. Dusseldorp,
Methodology and Statistics, Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK, Leiden, The Netherlands; e-mail: elise.dusseldorp@fsw.leidenuniv.nl

Optimal Random Projection Trees Ensemble

Nosheen Faiz, Adi Lausen, Metodi Metodiev, Zardad Khan, and Berthold Lausen

This paper develops the idea of creating an ensemble of accurate and diverse trees for improved classification accuracy: Optimal Random Projection Trees Ensemble (ORPTE), which is an extension of Optimal Trees Ensemble (OTE) [2]. Diversity in the base classification trees is introduced by the method of Random Projection as a dimensionality reduction tool that preserves pairwise distances between observations [2]. A sufficiently large number of trees is grown on bootstrap samples by using the Random Forest algorithm, each generated on a random projection of the training data. For maintaining accuracy in the base models, trees are ranked based on their out-of-bag error estimates and a certain proportion of the top ranked trees are selected. The selected trees are integrated as an ensemble for predicting new/unseen data. The proposed method was assessed on 25 benchmark datasets against seven competitor methods in addition to a simulation study. The results demonstrate that the proposed method outperformed its competitors in most of the data sets and the simulation setup.

Keywords: random projection, optimal trees, ensemble learning

References

1. Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., Lausen, B.: Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification **14**, 97–116 (2020)
2. Cannings, T.J., Samworth, R.J.: Random projection ensemble classification. Journal of the Royal Statistical Society B **79**(4), 1–38 (2017)

Nosheen Faiz
Abdul Wali Khan University, Pakistan, e-mail: nosheenfaiz09@gmail.com

Adi Lausen
Mathematical Sciences, University of Essex, UK, e-mail: a.lausen@essex.ac.uk

Metodi Metodiev
Life Sciences, University of Essex, UK, e-mail: mmetod@essex.ac.uk

Zardad Khan
Mathematical Sciences, University of Essex, UK, and Abdul Wali Khan University, Pakistan, e-mail: zardadkhan@awkum.edu.pk

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

Born-again and Bayesian Approaches for Improving the Performance of Decision Trees

Marjolein Fokkema

Breiman and Shang [1] proposed born-again trees, where a single decision is fit on a large artificially generated dataset $\{\mathbf{X}_{\text{gen}}, \mathbf{Y}_{\text{gen}}\}$. $\mathbf{X}_{\text{gen}}$ is constructed by resampling and permuting observations from the original training set of predictor variable values, and $\mathbf{Y}_{\text{gen}}$ are the predictions from a black-box method with high predictive accuracy.

The born-approach improves the predictive accuracy of single decision trees. Also, it provides a general approach for improving the predictive performance of inherently interpretable methods, as well as explaining the predictions of a black-box method using an interpretable method.

We present results on improvements and extensions of the born-again approach: We apply it to mixed-effects decision trees, and we employ the posterior predictive distribution of a Bayesian tree ensemble method (BART; [2]) to improve artificial data generation, by reducing the need for permutation and by incorporating uncertainty.

Keywords: decision trees, interpreble machine learning, born-again trees, bayesian additive regression trees

References

1. Breiman, L., & Shang, N. Born again trees. University of California, Berkeley, CA, Technical Report (1996).
2. Chipman, H.A., George, E.I., & McCulloch, R.E.: BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), 266-298 (2010).
3. Markovitch, B., & Fokkema, M.: Improved prediction rule ensembling through model-based data generation (2021). https://arxiv.org/abs/2109.13672

Marjolein Fokkema
Department of Methods & Statistics, Institute of Psychology, Wassenaarseweg 52, Leiden,
e-mail: m.fokkema@fsw.leidenuniv.nl

Evolution of Media Coverage on Climate Change and Environmental Awareness: an Analysis of Tweets from UK and US Newspapers

Gianpaolo Zammarchi, Maurizio Romano, and Claudio Conversano

Climate change represents one of the biggest challenges of our time. Newspapers might play an important role in raising awareness on this problem and its consequences. We collected all tweets posted by six UK and US newspapers in the last decade to assess whether 1) the space given to this topic has grown, 2) any breakpoint can be identified in the time series of tweets on climate change, and 3) any main topic can be identified in these tweets. Overall, the number of tweets posted on climate change increased for all newspapers during the last decade. Although a sharp decrease in 2020 was observed due to the pandemic, for most newspapers climate change coverage started to rise again in 2021. While different breakpoints were observed, for most newspapers 2019 was identified as a key year, which is plausible based on the coverage received by activities organized by the Fridays for Future movement. Finally, using different topic modeling approaches, we observed that, while unsupervised models partly capture relevant topics for climate change, such as the ones related to politics, consequences for health or pollution, semi-supervised models might be of help to reach higher informativeness of words assigned to the topics.

Keywords: climate change, twitter, environment, time series, topic modeling

Gianpaolo Zammarchi · Maurizio Romano · Claudio Conversano
University of Cagliari, Viale Sant'Ignazio 17, 09123, Cagliari,
e-mail: {gp.zammarchi,romano.maurizio,conversa}@unica.it

Improving Classification of Documents by Semi-supervised Clustering in a Semantic Space

Jasminka Dobša and Henk A.L. Kiers

In the paper we propose method for representation of documents in a semantic lower-dimensional space based on the modified Reduced k-means method which penalize clusterings that are distant from classification of train documents given by experts. Iterative method of the Reduced k-means (RKM) [1] enables simultaneously clustering of documents and extraction of factors. By projection of documents represented in the vector space model on extracted factors, documents are clustered in the semantic space in a semi-supervised way because clustering is guided by classification given by experts, which enables improvement of classification performance of test documents.
Classification performance is tested for classification by logistic regression and support vector machines (SVMs) for classes of Reuters-21578 data set. It is shown that representation of documents by the RKM method with penalization improves average precision of classification for 25 largest classes of Reuters collection for about 5,5% with the same level of average recall in comparison to the basic representation in the vector space model. In the case on classification by logistic regression, representation by the RKM with penalization improves average recall for about 1% in comparison to the basic representation.

Keywords: classification of textual documents, lsa, reduced k-means

References

1. De Soete, G., Carroll, J.D.: K-means clustering in a low-dimensional Euclidean space.. In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., Burtschy, B. (eds.) New Approaches in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 212-219. Springer, Heidelberg (1994)

Jasminka Dobša
Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 40000 Varaždin, Croatia, e-mail: jasminka.dobsa@foi.hr

Henk A.L. Kiers
Department of Psychology, University of Groningan, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, e-mail: h.a.l.kiers@rug.nl

Is It Hate or Criticism? An Exploratory Approach to Negative Comments on YouTube

Manuela Schmidt

As many communities on YouTube develop around controversial topics, marginalized identities, and extreme ideological positions, negative, anti-social or hateful commenting behavior on YouTube has garnered increasing interest. Whereas previous studies analyze these comments by sentiment analysis or use of profanity only, in this exploratory approach, the analysis of negative comments on YouTube is expanded on by combining sentiment analysis [1], a dictionary approach for defining and matching video topics using automatically generated transcripts, a dictionary matching for swear words [2], and video and comment metadata. A small group of content creators was selected for the analysis.

The combination of comment sentiment, the relation to a video and use of profanity allowed for the classification into *four* groups of different sizes: Impoliteness, Incivility, Flaming and Criticism [3].

Comparing approaches, sentiment analysis alone shows a similar prevalence of negative comments between creators. When the classification is applied, a different distribution of negative groups between the creators is observable along expectations. The frequency, commonality, and difference of the language used show that while the groups share multiple expressions, especially those carrying a negative sentiment value, they make distinct use of crucial terms. While this classification is valuable for gaining insight and nuance into negative comments, some applicability issues for larger studies will also be discussed.

Keywords: text mining, social science, social media, sentiment analysis

References

1. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, **63(1)**, 163-173 (2012)
2. Dumm, S., Niekler, A. Methoden, Qualitätssicherung und Forschungsdesign. In M. Lemke, G. Wiedemann (eds.), Text Mining in den Sozialwissenschaften, pp. 89–116. Springer VS, Wiesbaden (2015)
3. Barnes, R. Uncovering Online Commenting Culture: Trolls, Fanboys and Lurkers. Palgrave Macmillan, Cham (2018)

Manuela Schmidt
Institute of Sociology, Bonn University, Germany, e-mail: manuela-schmidt@uni-bonn.de

A Time-varying Text Based Ideal Point Model to Infer Partisanship in the U.S. Senate

Sourav Adhikari, Bettina Grün, and Paul Hofmarcher

Ideal point models analyze lawmakers' votes, speeches, press statements and social media posts to quantify their political positions along a latent continuum. In this work, we extend the text based ideal point model [1] to obtain a time-varying version to study the evolution of the ideological positions of lawmakers over time and assess the change in partisanship among representatives from two political parties. We aim to confirm recent findings regarding the increase in partisanship manifested in speeches by Republicans and Democrats in the U.S. Senate during the last years [2]. These findings were drawn using a penalized estimator for measuring group differences in choices with high-dimensional data and text analysis was based on manually pre-defined topics. By contrast, the time-varying text based ideal point model estimated using variational inference [3] infers topics in a data-driven way and does not use party membership to determine the ideological positions and thus is not susceptible to overrating spurious differences in vocabulary use of different party members. Drawing the same substantive conclusions based on the results of two different statistical text analysis methods provides evidence for their robustness.

Keywords: high-dimensional data, text mining, topic model, variational techniques

References

1. Vafa, K., Naidu, S., Blei, D.: Text based ideal points. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5345–5357 (2020)
2. Gentzkow, M., Shapiro, J.M., Taddy, M.: Measuring group differences in high-dimensional choices: Method and application to congressional speech. Econometrica. **87** 1307–1340 (2019)
3. Blei, D., Kucukelbir, D., McAuliffe, J. D.: Variational inference: A review for statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017)

Sourav Adhikari
Institute for Statistics and Mathematics, WU Wien, e-mail: sourav.adhikari@wu.ac.at

Bettina Grün
Institute for Statistics and Mathematics, WU Wien, e-mail: bettina.gruen@wu.ac.at

Paul Hofmarcher
Department of Economics, Paris Lodron University of Salzburg,
e-mail: paul.hofmarcher@plus.ac.at

Oracle-LSTM: a Neural Network Approach to Mixed Frequency Time Series Prediction

Alessandro Bitetto and Paola Cerchiello

In the context of macro-economic indicators there are two main concerns regarding the frequency of the variables. The first is related to MIxed DAta Sampling (MIDAS), i.e. some indicators are reported annually, some quarterly, other monthly. The second deals with the need of forecasting predictions between reporting dates, e.g. before the end of the year, and it is known as "nowcasting". Existing methods rely on the alignment of high-frequency input data to low-frequency target variable by the means of lagged variables and their temporal-decaying weighting. We develop a two-steps algorithm that makes use of two Recurrent Neural Networks. The first, called Oracle, is a Deep Autoregressive network and predicts the target variable at high-frequency given past information. The second, called Predictor, is Long-Short Term Memory (LSTM) network and learns the relationship between Oracle's predictions and high-frequency input data. The prediction error is a weighted average of two terms: one compares the observed low-frequency target with predictions of both Oracle and Predictor, the other compares the Predictor's high-frequency predictions with the Oracle's ones. Our model is tested on both simulated data, where we know the generated high-frequency data, and real macro-economic data. Our results show better performances compared to classical approach. Moreover, we apply gradient-based interpretability methods to estimate the input features' importance in the predictions.

Keywords: mixed frequency data, artificial neural networks, lstm, nowcast

Alessandro Bitetto
University of Pavia, Via San Felice al Monastero 5 - Pavia, Italy,
e-mail: alessandro.bitetto@unipv.it

Paola Cerchiello
University of Pavia, Via San Felice al Monastero 5 - Pavia, Italy,
e-mail: paola.cerchiello@unipv.it

Time Series of Counts Under Censoring

Isabel Silva, Maria Eduarda Silva, Isabel Pereira, and Brendan McCabe

Censored time series arise when explicit limits are placed on the observed data and occur in several fields including environmental monitoring, economics, medical and social sciences. The censoring may due to measuring device limitations, such as detection limits in air pollution or mineral concentration in water. Censoring may also occur when constraints or regulations are imposed, such as in international trade studies where exports and imports are subject to trade barriers or hours worked, often treated as censored variables. This work considers time series of counts under censoring, focusing on the Poisson first-order integer-valued autoregressive (PoINAR) models ([3] for details on PoINAR(1) models). This class, while being simple and flexible, is useful for modelling positive-valued and integer-valued time series possessing an autoregressive structure with non-negative serial correlation. We investigate two natural approaches to analyse censored PoINAR(1) time series under the Bayesian framework: the Approximate Bayesian Computation (ABC) methodology [2] and the Gibbs sampler with Data Augmentation (GDA) approach [1]. Both approches may also be valuable to analyse time series of counts with missing data.

Keywords: bayesian estimation, censored time series, poisson inar(1) model

References

1. Chib, S.: Bayes Inference in the Tobit Censored Regression Model. J. Econom. **51**, 79–99 (1992)
2. Plagnol V., Tavaré S.: Approximate Bayesian computation and MCMC. In: Niederreiter H (ed) Monte Carlo and quasi-Monte Carlo methods, pp. 99–113. Springer, Heidelberg (2004)
3. Scotto, M. G., Weiß, C. H., Gouveia, S.: Thinning-based models in the analysis of integer-valued time series: a review. Stat. Model. **15**, 590–618 (2015)

Isabel Silva
Faculdade de Engenharia, Universidade do Porto and CIDMA, Portugal, e-mail: ims@fe.up.pt

Maria Eduarda Silva
Faculdade de Economia, Universidade do Porto and LIAAD-INESC TEC, Portugal
e-mail: mesilva@fep.up.pt

Isabel Pereira
Departamento de Matemática, Universidade de Aveiro and CIDMA, Portugal,
e-mail: isabel.pereira@ua.pt

Brendan McCabe
Management School, University of Liverpool, UK, e-mail: Brendan.Mccabe@liverpool.ac.uk

Multivariate Time Series Feature Extraction via Multilayer Networks

Vanessa Freitas Silva, Maria Eduarda Silva, Pedro Ribeiro, and Fernando Silva

The extraction of features from time-indexed data has proved to be an important preliminary task in many applications of time series analysis, such as classification, clustering and forecasting. Finding a set of features that summarizes the main characteristics of such data is therefore a crucial task, which usually involves conventional statistical and non-linear measures of time series analysis [1]. Complementary, features based on complex network methods have been shown to be useful to characterize time series data [2, 3]. For multivariate time series (MTS) settings, feature extraction is even less trivial due to temporal and cross-dimension dependencies. The existing methods and models are often developed under certain constraints and for specific problems, and therefore new methodological and computational tools are required.

Multilayer networks are complete structures able of mapping the internal and external temporal dependencies of MTS through intra and inter-network connections [2]. In this work, we introduce novel MTS features based on a new multilayer visibility network mapping method. To demonstrate its applicability, we performed a MTS clustering task based on the proposed features set.

Keywords: multivariate, time series, visibility graphs, feature extraction

Acknowledgements The authors gratefully acknowledge support from FCT (SFRH/BD/139630/2018) and PREFERENTIAL (PTDC/MAT-STA/28243/2017).

References

1. Montero-Manso, P., Athanasopoulos, G., Hyndman, R.J., Talagala, T.S.: FFORMA: Feature-based forecast model averaging. Int. J. Forecast. **36**, 86–92 (2020)
2. Silva, V.F., Silva, M.E., Ribeiro, P., Silva, F.: Time series analysis via network science: Concepts and algorithms. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **11**, e1404 (2021)
3. Silva, V.F., Silva, M.E., Ribeiro, P., Silva, F.: Novel features for time series analysis: a complex networks approach. Data Min. Knowl. Disc. (2022)

Vanessa Freitas Silva
INESC TEC/CRACS, FCUP, University of Porto, Portugal, e-mail: vanessa.silva@fc.up.pt,

Maria Eduarda Silva
INESC TEC/LIAAD, FEP, University of Porto, Portugal, e-mail: mesilva@fep.up.pt,

Pedro Ribeiro
INESC TEC/CRACS, FCUP, University of Porto, Portugal, e-mail: pribeiro@fc.up.pt

Fernando Silva
INESC TEC/CRACS, FCUP, University of Porto, Portugal, e-mail: fmsilva@fc.up.pt

On the Use of the Choquet Fuzzy Integral to Aggregate Predictions of Time Series: an Application to Economic (and Other Types of) Data

Diogo Alves, José Matos, and Sandra Silva

The Choquet Integral with respect to a given fuzzy measure is a powerful aggregation operator that fuses several sources of information into a single value [1]. An as of yet unexplored application is the use of the method to aggregate predictions of time series (as well as supervised learning representations) methods. Time series have an internal structure based on temporal ordering of the data, and the fact that changes in the relative ordering of observations would change the meaning of the data is a fact that presents its own set of challenges in the context of Machine Learning (see [2, 3] for examples of applications). The objective is twofold: first, to compare the methods used in terms of performance. Second, to obtain a model acting as an ensemble, able to muster the relative strengths of each method into a more accurate (in the sense of smaller in and out-of-sample errors with respect to its components) super predictor, in a "wisdom of crowds" effect. An application example with 8 time series (of both synthetic and "real-life" origin) is explored.

Keywords: fuzzy measures, Choquet integral, time series forecasting, arima, supervised learning

References

1. Grabisch, M., Labreuche, C.: Fuzzy Measures and Integrals in MCDA. In: Greco, S., Ehrgott, M., Figueira, J. (eds) Multiple Criteria Decision Analysis. International Series in Operations Research & Management Science, vol **233**, pp 553-603. Springer, New York, NY (2016)
2. Ribeiro, M.H.D.M., Coelho, L.S.: Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. App. Soft Comp. **86**, 105837 (2020)
3. Cicceri, G., Inserra, G. and Limosani, M. A machine learning approach to forecast economic recessions—an italian case study. Mathematics **8(2)**, 241 (2020)

Diogo Alves
Whiteshield Partners, e-mail: diogo.alves@whiteshield.com

José Matos
Centre of Mathematics and Faculty of Economics, University of Porto,
e-mail: jamatos@fep.up.pt

Sandra Silva
Faculty of Economics, University of Porto, e-mail: sandras@fep.up.pt

Some Biplot Alternatives

Patrick Groenen

Biplots provide a valuable tool for exploring and visualizing the relations between two entities, often individuals and attributes. They originate from principle components analysis, but have been applied to many other techniques that provide a decomposition between. The most well known biplot provides a projection interpretation: attributes are shown as vectors in a low-dimensional space, individuals as points, and the projection of an individual onto the direction of the variable is proportional to the reconstructed data value of this individual for that attribute. Here we discuss two alternatives to the standard biplot. The first one is the so-called area biplot [1] that can be used as an alternative to every standard projection biplot. Its main difference is that the estimate of the data is given by the area formed by the origin and two points. The second variety is the nonlinear biplot with a distance interpretation: the reconstructed value on a variable of each sample point is obtained by finding the nearest marker point on a nonlinear curve representing the variable [2]. Each type of biplot will be explained briefly and compared with the standard biplot.

Keywords: biplot, visualisation

References

1. Gower, J. C., Groenen, P. J. F., van de Velden, M.: Area biplots. J Comput Graph Stat, **19** (1), 46-61 (2010)
2. Groenen, P. J. F., Le Roux, N. J., Gardner-Lubbe, S.: Spline-based nonlinear biplots. Adv Data Anal Classif, **9** (2), 219–238 (2015)

Patrick Groenen
Econometric Institute, Erasmus University Rotterdam, the Netherlands,
e-mail: groenen@ese.eur.nl

Biplots in Dimension Reduction and Clustering

Alfonso Iodice D'Enza, Angelos Markos, and Michel van de Velden

In unsupervised learning, dimension reduction (e.g., PCA) and distance-based clustering are often applied sequentially: the distances used to cluster the observations are computed on the reduced dimensions. Since the dimension reduction step does not take into account the possible cluster structure, it is possibly detrimental to the clustering step. Methods for joint dimension reduction and clustering combine the two in a single optimization problem which is solved using iterative procedures alternating the two steps. Just like for principal component methods, different approaches have been proposed that deal with continuous, categorical or mixed-type data. In particular, for continuous data, reduced K-means [1] combines principal component analysis with K-means clustering; for categorical data, cluster correspondence analysis [2] combines correspondence analysis with K-means; for mixed-type data, mixed Reduced K-means [3] combines factor analysis for mixed data with K-means. The biplot visualization of the solution is of particular interest for interpretation purposes: in fact, the low-dimensional map can be very helpful for cluster characterization. In this work, we illustrate the use of biplots in the context of dimension reduction and clustering.

Keywords: biplot, dimension reduction, clustering

References

1. De Soete, G., Carroll, J. D.: K-means clustering in a low-dimensional Euclidean space. In: E. Diday, et al. (eds.), New approaches in classification and data analysis, pp. 212–219, Springer, Heidelberg (1994)
2. van de Velden, M., Iodice D'Enza, A., Palumbo, F.: Cluster correspondence analysis. Psychometrika **82**(1), 158–185 (2017)
3. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics **11**(3), e1456 (2019)

Alfonso Iodice D'Enza
University of Naples Federico II, Via L. Rodino, 22, Naples, Italy, e-mail: iodice@unina.it

Angelos Markos
Democritus University of Thrace, Nea Hili, GR-68100, Alexandroupoli, Greece,
e-mail: amarkos@eled.duth.gr

Michel van de Velden
Erasmus University Rotterdam, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands, e-mail: vandevelden@ese.eur.nl

Biplots: a Sophisticated Multivariate Approach or a User-Friendly Technique?

Manuel Rui Alves

Predictive biplots are based on the visualization of the results of multivariate analyses through the relations between the objects and the original variables equipped with measuring scales, avoiding the interpretation of latent variables, making them a useful tool for non-statisticians. However, to achieve this desideratum, biplots need to be automatically simplified as is commonly done with any multivariate analysis.

Through the definition of an axis mean standard predictive error (mspe), which evaluates the error that is made by the analyst when reading a biplot axis, it is possible to automatically select the axes to be included in the biplots, to define the number of dimensions necessary to conveniently describe any given problem, to enable the evaluation of outliers and avoid common overestimations. A series of "AutoBiplot" functions have been written in R to produce PCA [1] and CCA [2] biplots and can be easily adjusted to many other multivariate analyses [3]. Apparently forgotten, interpolative biplots can be used in laboratory practical work, and can be automatically produced following a similar strategy, based on the definition of an overall standard interpolative error (osie).

Although the techniques for the automatic production of biplots are available, a wide use of biplots is still restricted mainly to statisticians. Therefore, a good way to commemorate fifty years of biplots may be to envisage ways of rendering these methods available to any person needing to apply multivariate analysis.

Keywords: biplots, multivariate, autobiplots

References

1. Alves, M. R.: Evaluation of the predictive power of biplot axes to automate the construction and layout of biplots based on the accuracy of direct readings from common outputs of multivariate analyses: 1. application to principal component analysis. J. Chem., **26**(5), 180–190 (2012)
2. Alves, M.R.: Getting full control of canonical correlation analysis with the AutoBiplot.CCA function. AIP Conference Proceedings 1738, 370015 (2016);
3. Barbosa, C., Oliveira, M. B., Alves, M. R.: Chemometrics in food authentication. In: Oliveira, B., Mafra, I., Amaral, J. (eds) Current topics on food authentication, pp. 237-268. Transworld Research Network, Kerala, India (2011).

Manuel Rui Alves
Center for Research and Development in Agrifood Systems and Sustainability, Viana do Castelo, Portugal, e-mail: mruialves@estg.ipvc.pt

PLS-based Principal Balances for Regression and Classification with High-dimensional Compositional Data

Viktorie Nesrstová, Ines Wilms, Karel Hron, Josep A. Martín-Fernández, Peter Filzmoser, and Javier Palarea-Albaladejo

Compositional data naturally occur in many different research fields, such as geochemistry or metabolomics. Due to their relative nature, compositions cannot be directly analysed using standard statistical methods. Instead, it is common to express compositional data as log-ratio coordinates with respect to an orthonormal basis of their sample space, see [1, 2], named as orthonormal logratio (olr) coordinates. Principal balances (PBs) are a specific class of olr coordinates, which are a suitable choice in a high-dimensional context, see [3]. They are constructed such that the first few coordinates capture most of the original data variability. In this work, we adapt the PB procedure introduced in [3] to be used for variable selection in regression and classification problems with a high-dimensional composition acting in an explanatory role. Moreover, hereby we extend popular logcontrast models to consider other potentially interesting (orthonormal) logcontrasts. For this, partial least squares (PLS) estimation is embedded into the construction of the PBs. The performance of the proposal is demonstrated using simulated and real-world data.

Keywords: high-dimensional compositional data, principal balances, PLS regression and classification

References

1. Filzmoser, P., Hron, K., Templ, M.: Applied compositional data analysis. Springer, Cham (2018)
2. Pawlowsky-Glahn, V., Egozcue, J.J, Tolosana-Delgado, R.: Modeling and analysis of compositional data. John Wiley & Sons (2015)
3. Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R.: Advances in principal balances for compositional data. Math. Geosci. **50**, 273–298 (2018)

V. Nesrstová, K. Hron
Palacký University Olomouc, the Czech Republic,
e-mail: viktorie.nesrstova@gmail.com,karel.hron@upol.cz

I. Wilms
Maastricht University, Maastricht, The Netherlands,
e-mail: i.wilms@maastrichtuniversity.nl

J. A. Martín-Fernández, J. Palarea-Albaladejo
University of Girona, Girona, Spain,
e-mail: josepantoni.martin@udg.edu, javier.palarea@udg.edu

P. Filzmoser
Vienna University of Technology, Vienna, Austria, e-mail: peter.filzmoser@tuwien.ac.at

Clustering Count Data Using Compositional Methods

Marc Comas-Cufí, Josep A. Martín-Fernández, Glària Mateu-Figueras, and Javier Palarea-Albaladejo

Multivariate count data are multivariate vectors of non-negative integers. When the total number of counts depends on varying external factors (e.g. duration of the counting process or ability to measure a count), the relative magnitude of the observed values is of special importance. These data are called point-counting data in some applied fields, and they are affected by compositional variability and multinomial counting uncertainties [4]. For clustering analysis purposes, it is crucial to consider both sources of variability. Compositional variability is commonly modeled using the Dirichlet or the logratio-normal distribution, leading respectively to the classical Dirichlet-multinomial distribution and the logratio-normal-multinomial (LRNM) distribution [2]. In model-based clustering, these models are usually included as components in a finite mixture model [1,3].

In this contribution, we propose a fast approach for clustering analysis of point-counting data based on the LRNM model. A part of the procedure takes care of the compositional aspect through logratio coordinates, including the treatment of zero counts as necessary. Moreover, through the computation of the posterior distribution conditioned to the measured variability and the observed data, the multinomial variability is accounted for by using simulated samples of the latent compositional process. These samples are combined into a final partition of the original sample by using clustering ensembling methods. We will illustrate our proposal using different datasets.

Keywords: multivariate count data, clustering, compositional data analysis

References

1. Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G. and Palarea-Albaladejo, J.: Model-based clustering of count data based on the logistic-normal-multinomial distribution. 2017 International Federation of Classification Societies Conference, Tokyo, Japan (2017)
2. Comas-Cufí, M., Martín-Fernández, J.A., Mateu-Figueras, G. and Palarea-Albaladejo, J.: Modelling count data using the logratio-normal-multinomial distribution. SORT **44(1)**, 99–126 (2020)
3. Fang, Y. and Subedi, S.: Clustering microbiome data using mixtures of logistic normal multinomial models. arXiv preprint: 2011.06682 (2020)
4. Vermeesch, P.: Statistical models for point-counting data. Earth and Planetary Science Letters **501**, 112–118 (2018)

Marc Comas-Cufí
University of Girona, e-mail: marc.comas@udg.edu

Urban Development Paths in Poland: Multidimensional Perspective

Barbara Batóg and Jacek Batóg

The development of countries and regions is strongly dependent on the socio-economic situation characterising the largest urban centres [3]. This is caused not only by a higher intensity of processes taking place in urban space, but also by the predominance of the population living in cities in comparison to rural areas. The increasing role of urban systems in the creation of domestic product and the accumulation of services and innovations on their area results in a growing interest in studying the phenomenon of urbanisation and concentration of social and economic activities realised in urban space [1]. The main objective of the study will be the construction of development paths of the largest Polish cities, which are the capitals of a given region, in 2004-2020. To achieve it, the authors will use their own methodological proposal from the area of multidimensional analysis. An additional aspect of the study will be a comparison of the similarity of the changes and levels of development of particular cities with the use of selected measures of time series similarity and cluster analysis [2]. The analyses are to answer the following research questions: (i) what was the socio-economic development of Polish voivodship cities characterised by in the last years, (ii) has this development proceeded in a similar way for all analysed cities, and (iii) what has been the impact of the last two crises on the dynamics and heterogeneity of development of the analysed cities?

Keywords: cities, development, economic crisis, multivariate analysis, temporal analysis

References

1. Bogdański M.: Socio-economic potential of Polish cities – a regional dimension, Bull. Geogr. Socio. Econ. Ser. **17**, 13-20 (2012)
2. Fanni Z., Khakpour B.A., and Heydari A.: Evaluating the regional development of border cities by TOPSIS model (case study: Sistan and Baluchistan Province, Iran). Sustain. Cities Soc. **10**, 80-86 (2014)
3. Zoeteman K., Mommaas H., and Dagevos J.: Are larger cities more sustainable? Lessons from integrated sustainability monitoring in 403 Dutch municipalities. Environ. Dev. **17**, 57-72 (2016)

Barbara Batóg
University of Szczecin, Mickiewicza 64 71-101 Szczecin, e-mail: barbara.batog@usz.edu.pl

Jacek Batóg
University of Szczecin, Mickiewicza 64 71-101 Szczecin, e-mail: jacek.batog@usz.edu.pl

Analyzing the Evolution of EU Countries and Indicators of Europe 2020 Agenda

Adelaide Figueiredo and Fernanda Figueiredo

In this study we analyze the evolution of the European Union countries and of some indicators of Europa 2020 agenda in the following areas: Employment, Education, Research and Development, Poverty and Social Exclusion, and Climate Change and Energy. More precisely, we collected data from Pordata during the period 2010-2019 on the following indicators: employment rate; early leavers from education and training rate; population, aged 30 to 34, with higher education; expenditure on R&D as % of GDP; population at risk of poverty; greenhouse gas emissions; renewable energy consumption; primary energy consumption; final energy consumption.
We start with a preliminary data analysis, and some countries appear as outliers in some of the variables and for some years of the period, which would be expected. However, we highlight that Luxembourg is a severe outlier and had to be discarded from the analysis to allow a better comparison and differentiation of the other countries. Additionally, as we do not have all the target values for the United Kingdom and for the variable population at risk of poverty, we did not consider this country and variable in the further multivariate analysis. We applied the Statis methodology, developed in [1] and [2], to analyze the evolution of the European countries and of the indicators referred above, during the period 2010-2019. The trajectories of the countries and of the variables under study along the period 2010-2019 help us to understand how the Europe 2020 strategy is being achieved.

Keywords: Europe 2020, European countries, Statis methodology

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects LA/P/0063/2020 (INESC TEC) and UIDB/00006/2020 (CEAUL).

References

1. Lavit, C.: Analyse Conjointe de Tableaux Quantitatives. Masson (1988)
2. Lavit, C., Escoufier, Y., Sabatier, R., Traissac, P.: The Act (Statis Method). Comput. Statist. and Data Analysis **18**, 87–119 (1994)

Adelaide Figueiredo
Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC, Rua Dr Roberto Frias, 4200-464 Porto, Portugal, e-mail: adelaide@fep.up.pt

Fernanda Figueiredo
Faculdade de Economia da Universidade do Porto and CEAUL, Rua Dr Roberto Frias, 4200-464 Porto, Portugal, e-mail: otilia@fep.up.pt

Google Trends as a Macroeconomic Predictor: Behind the Scenes

Eduardo Andre Costa and Maria Eduarda Silva

Data sourced from online activities and particularly from Google Trends (GT) has been gaining importance in the literature as predictors for economic indicators. Recent research evidence relationships between GT and a range of outcomes, thus establishing GT as a complementary data source. GT yields a measure of the interest of Google search-engine users in a subject or a specific keyword over time, resulting in a normalised index [1]. The benefits of GT as a data supplier include its promptness, inexpensive collection costs, mixed sampling frequency and approaches to diversified research areas. There are, however, two issues associated with GT that require attention. The first regards the frequency of the returned index, which depends on the time length and span required, leading to limited historical data in high-frequency sampling. If daily data are necessary, then GT is limited to 9-months of data, whereas monthly sampling grants more than 5-years of data. Since GT normalises indexes, stacking multiple time frames to obtain extended periods of high-frequency data would conceal eventual trends in the underlying data; thus, temporal disaggregation procedures must be applied [2]. The second issue involves the data source sampling noise since different indexes are produced on distinct days, even when all the other constraints (time length, time span, subject and keywords) are kept constant. As GT considers aggregated searches based on samples, performing collections over multiple days and summarising them into a single measure controls for the data uncertainty [1]. This work addresses the construction of a time series predictor from GT's indexes in a nowcasting exercise with mixed frequency data.

Keywords: google trends, data source, high-frequency sampling

Acknowledgements Eduardo Andre Costa gratefully acknowledges support from CEF.UP/FCT (UIDB/04105/2020) and FCT (2021.07583.BD).

References

1. McLaren, N., Shanbhogue, R. Using internet search data as economic indicators. Bank Engl. Q. Bull. **51(2)**, 134–140 (2011)
2. Eichenauer, V.Z., Indergand, R., Martínez, I.Z., Sax, C. Obtaining consistent time series from Google Trends. Econ. Inq. **60(2)**, 694–705 (2022)

Eduardo Andre Costa
Faculdade de Economia, Universidade do Porto, e-mail: ecosta.phd@fep.up.pt

Maria Eduarda Silva
INESC TEC/LIAAD, Faculdade de Economia, Universidade do Porto,
e-mail: mesilva@fep.up.pt

COVID-19 Pandemic: a Methodological Model for the Analysis of Government Preventing Measures and Health Data Records

Theodore Chadjipadelis and Sofia Magopoulou

The study aims to investigate the associations between the government's response measures during the COVID-19 pandemic and weekly incidence data (positivity rate, mortality rate and testing rate) in Greece. The study focuses on the period from the detection of the first case in the country (26th February 2020) to the first week of 2022 (08th January 2022). Data analysis was based on Correspondence Analysis on a fuzzy-coded contingency table, followed by Hierarchical Cluster Analysis (HCA) on the factor scores. Results revealed distinct time periods during which interesting interactions took place between control measures and incidence data.

Keywords: hierarchical cluster analysis, correspondence analysis, covid-19, evidence-based policy making

References

1. Asan, Z., Greenacre, M.: Biplots of fuzzy coded data. Fuzzy Sets and Systems, **183**(1), 57-71 (2011)
2. Chadjipadelis, T.: Facebook profile, https://www.facebook.com/theodore.chadjipadelis (2022)
3. Benzécri, J.P.: L'Analyse des Donnés. 2. L'Analyse des Correspondances. Dunod, Paris (1973)
4. Hale, T., Petherick, A., Phillips, T., and Webster, S.: Variation in government responses to COVID-19. Blavatnik school of government working paper, 31, 2020-11 (2020)
5. Karapistolis, D.: Software Method of Data Analysis MAD. (2010) http://www.pylimad.gr/
6. Markos A., Moschidis O., Chadjipadelis T.: Hierarchical clustering of mixed-type data based on barycentric coding (2022) https://arxiv.org/submit/4142768
7. Moschidis O., Chadjipadelis T.: A method for transforming ordinal variables. In: Palumbo F., Montanari A., Vichi M. (eds) Data Science. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham. https://doi.org/10.1007/978-3-319-55723-6_22 (2017)
8. Papadimitriou, G., Florou, G.: Contribution of the Euclidean and chi-square metrics to determining the most ideal clustering in ascending hierarchy (in Greek). In Annals in Honor of Professor I. Liakis, 546-581. University of Macedonia, Thessaloniki (1996)

Theodore Chadjipadelis · Sofia Magopoulou
Aristotle University of Thessaloniki,
e-mail: chadji@polsci.auth.gr;sofimago@polsci.auth.gr

Detecting Fabricated Interviews Using the Hamming Distance

Joerg Blasius

In the research literature on survey methodology, there is considerable discussion of interviewer effects and how to prevent data fabrication; however, there is little discussion on the detection of data fabrication by interviewers in published data, and there are even fewer papers examining the phenomenon of employees of survey research organizations fabricating data. Among them, Blasius and Thiessen ([1]) show for the PISA 2009 principal data that employees of survey research organizations in some countries duplicate cases to generate data. While the authors focus on exact copies, more sophisticated data fabrication techniques might include duplicating whole cases and changing a few entries afterwards. By calculating Hamming distances and applying them to the same data, we show that - in some countries in particular - large parts of the data have been duplicated, and most of them have been retrospectively modified to a small degree.

Keywords: fabricated data, string distances, pisa data

References

1. Blasius, J., Thiessen, V.: Should we trust survey data? Assessing response simplification and data fabrication. Soc.Sci.Res. **52**, 479–493 (2015)

Joerg Blasius
Institute of Political Science and Sociology, Bonn University, Germany
e-mail: jblasius@uni-bonn.de

Digital Development and Internet Use in the European Union Countries

Fernanda Figueiredo and Adelaide Figueiredo

In this study we analyze how people from the European Union countries are prepared to work and use in their lives the digital technologies. It is also interesting to know whether the least developed countries are close or quite distant from the others in this way to the digital.

We considered some variables associated with digital skills, digital economy and digital society, and collected the data from Eurostat database during the period 2010-2020. To analyze the data, we applied a Double Principal Component Analysis (DPCA), a method of multivariate data analysis introduced in Bouroche [1] to analyze three-way data with quantitative variables. We considered six different data tables, corresponding to the years 2010, 2012, 2014, 2016, 2018 and 2020, with the same countries and variables. As the UK does not belong to the EU countries in 2020, and all the tables must have the same countries to apply a DPCA, we considered the values obtained in 2019 for the UK to include it in the analysis.

The results allow to identify the differences and similarities between countries and variables along the period of time 2010-2020, more precisely, to study the evolution trends of the countries and the evolution of the relations between the different variables.

Keywords: digital, double principal component analysis, principal component analysis

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects UIDB/00006/2020 (CEAUL) and LA/P/0063/2020 (INESC TEC).

References

1. Bouroche, J.M.: Analyse des donnés ternaires: la double analyse en composantes principales. Thèse de 3ème cycle. Université de Paris VI (1975)

Fernanda Figueiredo
Faculdade de Economia da Universidade do Porto and CEAUL, Rua Dr Roberto Frias, 4200-464 Porto, Portugal, e-mail: otilia@fep.up.pt

Adelaide Figueiredo
Faculdade de Economia da Universidade do Porto and LIAAD - INESC TEC, Rua Dr Roberto Frias, 4200-464 Porto, Portugal, e-mail: adelaide@fep.up.pt

Probabilistic Clustering with Local Alignement of Italian COVID-19 Death Curves

Marzia A. Cremona, Tobia Boschi, and Francesca Chiaromonte

Italy is one of the most hardly hit countries in the world by the COVID-19 pandemic. A striking aspect of the pandemic in Italy has been its heterogeneity. Indeed, Italian regions were hit at different times and with different strengths, especially during the first wave. We consider official COVID-19 death curves, as well as excess mortality curves for the 20 Italian regions. The goal is to cluster these misaligned functional data, in order to assess whether there are regions sharing similar pandemic patterns [1]. Importantly, we are looking for clusters based on a local similarity among curves, since patterns might differ only on a (misaligned) portion of the domain.

We develop probabilistic K-mean with local alignment (*probKMA*), a new functional data analysis method to locally cluster a set of curves and discover functional motifs, i.e. typical "shapes" that may recur several times along and across the curves capturing important local characteristics of these curves [2]. Using probKMA as a probabilistic clustering method to group COVID-19 curves we find two starkly different first waves of COVID-19 pandemics; an "exponential" one unfolding in Lombardia and the worst-hit areas of the north, and a milder, "flat(tened)" one in the rest of the country.Local alignments of curves provide an indication of the lags between different regions, which can be employed in subsequent analyses to associate patterns of mortality with functional covariates such as mobility and positivity [1].

Keywords: functional data analysis, local clustering, covid-19

References

1. Boschi, T., Di Iorio, J., Testa, L., Cremona, M.A., Chiaromonte F.: Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy. Scientific Reports **11**:17054 (2021)
2. Cremona, M.A., Chiaromonte F.: Probabilistic K-mean with local alignment for clustering and motif discovery in functional data. arXiv 1808.04773 (2020)

Marzia A. Cremona
Dept. of Operations and Decision Systems, Université Laval and CHU de Quebec – Université Laval Research Center Québec, Canada, e-mail: marzia.cremona@fsa.ulaval.ca

Tobia Boschi
Dept. of Statistics, Penn State University, University Park, USA, e-mail: tub37@psu.edu

Francesca Chiaromonte
Dept. of Statistics, Penn State University, University Park, USA and EMbeDS, Sant'Anna School of Advanced Studies, Pisa, Italy, e-mail: fxc11@psu.edu

Model Free Predictive Inference for Functional Kriging Techniques Based on Conformal Prediction

Andrea Diana, Elvira Romano, and Jorge Mateu

In the last years several geostatistical predictive techniques like universal kriging, kriging with external drift and kriging with residuals has been extended to the functional framework to predict curve at unmonitored location. Here we present a new approach for model-free spatial prediction for kriging methods based on the conformal prediction. We propose non conformity measures for this class of predictive methods and introduce a local spatial conformal prediction algorithm that yields valid functional prediction intervals without any distributional assumption. Practical solutions are provided to construct conformal intervals and are compared with existing validation methods. The approach is illustrated on some well known data sets, on simulated and on a real data.

Keywords: functional data modelling, prediction, model validation, conformal prediction

References

1. Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statist. Sinica **21**, (2011)
2. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal Prediction Bands for Multivariate Functional Data. Journal of Multivariate Data Analysis (2022).
3. Franco-Villoria, M., Ignaccolo, R.: Bootstrap based uncertainty bands for prediction in functional kriging, Spatial Statistics, Volume 21, Part A,pp. 130-148, (2017).
4. Franco-Villoria, M. Ignaccolo, R.: Universal, Residual, and External Drift Functional Kriging. In: Geostatistical Functional Data Analysis. Giraldo, R., Mateu, J. (eds.) Wiley & Sons Ltd, pp. 55-72, (2022).
5. Giraldo, R., Delicado, P. and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, **18**, 411-426.
6. Ignaccolo, R., Mateu, J. and Giraldo, R. :Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment*, 28, 1171-1186, (2014). .

Andrea Diana
Department of Mathematics and Physics, Universitá della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: andrea.diana@unicampania.it

Elvira Romano
Department of Mathematics and Physics, Universitá della Campania "Luigi Vanvitelli", Caserta, Italy, e-mail: elvira.romano@unicampania.it

Jorge Mateu
Department of Mathematics, Jaume I University, Castellòn, Spain, e-mail: mateu@mat.uji.es

Density Modelling via Functional Data Analysis

Stefano A. Gattone and Tonio Di Battista

Recent technological advances have eased the collection of big amounts of data in many research field. In this scenario an useful statistical technique is density estimation which represents an important source of information. One dimensional density functions represent a special case of functional data subject to the constraints to be non-negative and with a constant integral equal to one [1]. Because of these constraints, densities functions do not form a vector space and a naive application of functional data analysis (FDA) methods may lead to non valid estimates. To address this issue two main strategies can be found in the literature. In the first, the probability density functions (pdfs) are mapped into a linear functional space through a suitably chosen transformation [2]. Established methods for Hilbert space valued data can be applied to the transformed functions and the results are moved back into the density space by means of the inverse transformation. In the second strategy, pdfs are treated as an infinite dimensional compositional data since they are part of some whole which only carry relative information. An approach based on the Aitchison geometry for compositional data has been sketched in [3, 4]. In this work, by means of a suitable transformation, densities are embedded in the Hilbert space of square integrable functions where standard FDA methodologies can be applied.

Keywords: constrained estimator, functional data analysis, probability density functions.

References

1. Ramsay, J.O. and Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)
2. Petersen, A. and Muller, H.: Functional data analysis for density functions by transformation to a Hilbert space. Ann. Stat. **32**(1), 183–218 (2016)
3. Delicado, P.: Dimensionality reduction when data are density functions. Comput. Stat. Data. Anal. **55**, 401–420 (2011)
4. Hron, K., Menafoglio, M., Templ, M., Hruzova, K. and Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Comput. Stat. Data. Anal. **94**, 330–350 (2016)

Stefano A. Gattone
DISFIPEQ, University G. d'Annunzio of Chieti-Pescara, Italy, e-mail: gattone@unich.it

Tonio Di Battista DISFIPEQ, University G. d'Annunzio of Chieti-Pescara, Italy,
e-mail: gattone@unich.it

On Parsimonious Modelling via Matrix-variate t Mixtures

Salvatore D. Tomarchio

Mixture models for matrix-variate data have becoming more and more popular in the most recent years. One issue of these models is the potentially high number of parameters. To address this concern, parsimonious mixtures of matrix-variate normal distributions have been recently introduced in the literature. However, when data contains groups of observations with longer-than-normal tails or atypical observations, the use of the matrix-variate normal distribution for the mixture components may affect the fitting of the resulting model. Therefore, we consider a more robust approach based on the matrix-variate t distribution for modeling the mixture components. To introduce parsimony, we use the eigen-decomposition of the components scale matrices and we allow the degrees of freedom to be equal across groups. This produces a family of 196 parsimonious matrix-variate t mixture models. Parameter estimation is obtained by using an AECM algorithm. The use of our parsimonious models is illustrated via a real data application, where parsimonious matrix-variate normal mixtures are also fitted for comparison purposes.

Keywords: matrix-variate, mixture models, clustering, parsimonious models

Salvatore D. Tomarchio
University of Catania, Department of Economics and Business, Catania, Italy
e-mail: daniele.tomarchio@unict.it

Four Skewed Tensor Variate Distributions

Michael P.B. Gallaugher, Peter A. Tait, and Paul D. McNicholas

In recent years, data has become more and more complex, coming in many different forms. One such example is data that come in the form of higher order tensors. Some examples of these type of data are coloured images, video clips, and medical data. Just like in the multivariate and matrix variate cases, the tensor or multilinear normal distribution is most commonly used in the literature; however, in the area of clustering and classification, if the data is skewed or contain outliers, then the use of a tensor normal distribution may result in over fitting the number of groups. We will introduce four skewed tensor variate distributions which will be utilized for model-based clustering and classification. Parameter estimation and properties will be discussed, and simulated and real data will be used for illustration.

Keywords: high-order data, tensors, skewed distributions

Michael P. B. Gallaugher
Baylor University, Waco Texas, e-mail: Michael_Gallaugher@baylor.edu,

Peter A. Tait
McMaster University, Hamilton Ontario, e-mail: taitpa@mcmaster.ca

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

A Family of Skewed Power Exponential Mixture Models for Clustering and Classification

Utkarsh J. Dang, Michael P. B. Gallaugher, Ryan P. Browne, and Paul D. McNicholas

In model-based clustering, mixture models that can deal with varying cluster tail-weight, skewness, concentration, and kurtosis are increasingly becoming common. Mixtures of multivariate power exponential (MPE) distributions were previously shown to be competitive for clustering in comparison to other elliptical mixture distributions [1]. Here, we introduce a novel formulation of a multivariate skewed power exponential distribution and mixtures thereof to combine the flexibility of the MPE distribution with the ability to model cluster-specific skewness. These mixtures are more robust to departures from normality and can model skewness, varying tail weight, and peakedness within clusters. A family of parsimonious models is proposed using an eigen-decomposition of the scale matrix. For parameter estimation, a generalized expectation-maximization approach combining minorization-maximization and optimization based on accelerated line search algorithms on the Stiefel manifold is utilized. These mixtures are implemented both in the model-based clustering and classification frameworks. We illustrate performance on toy and benchmark data in a wide range of scenarios.

Keywords: model-based clustering, multivariate skewed power exponential, mixture models, classification

References

1. Dang, U.J., Browne, R.P., McNicholas, P.D.: Mixtures of multivariate power exponential distributions. Biometrics. **71**, 1081–1089 (2015)

Utkarsh J. Dang
Carleton University, 1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada,
e-mail: utkarsh.dang@carleton.ca

Michael P. B. Gallaugher
Baylor University, One Bear Place #97140 Waco, TX, 76798,
e-mail: michael_gallaugher@baylor.edu

Ryan P. Browne
University of Waterloo, 200 University Avenue West Waterloo, ON, Canada N2L 3G1,
e-mail: ryan.browne@uwaterloo.ca

Paul D. McNicholas
McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 3L8,
e-mail: paulmc@mcmaster.ca

Generating Collective Counterfactual Explanations in Score-based Classification via Mathematical Optimization

Jasone Ramírez-Ayerbe, Emilio Carrizosa, and Dolores Romero Morales

Due to the increasing use of Machine Learning models in high stakes decision-making settings, it has become increasingly important to be able to understand how models arrive at decisions. Assuming an already trained Supervised Classification model, an effective class of post-hoc explanations are counterfactual explanations [2], i.e., a set of actions that can be taken by an instance such that the given Machine Learning model would have classified it in a different class. In this talk, for score-based multiclass classification models, we propose novel Mathematical Optimization formulations to construct the so-called collective counterfactual explanations, i.e., explanations for a group of instances in which we minimize the perturbation in the data (at the individual and group level) to have them labelled by the classifier in a given group [1]. Although the approach is valid for any classification model based on scores, we focus on additive tree models, like random forests or XGBoost. Our approach is capable of generating diverse, sparse, plausible and actionable collective counterfactuals. Real-world data are used to illustrate our method.

Keywords: collective counterfactual explanations, score-based classification, mathematical optimization

References

1. Carrizosa, E., Ramírez-Ayerbe, J., Romero Morales , D.: Generating Counterfactual Explanations in Score-Based Classification via Mathematical Optimization. Technical Report IMUS, Sevilla, Spain (2022) doi: 10.13140/RG.2.2.22996.12168/1
2. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology, **31** 841 (2017)

Jasone Ramírez-Ayerbe
Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain, e-mail: mrayerbe@us.es

Emilio Carrizosa
Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain, e-mail: ecarrizosa@us.es

Dolores Romero Morales
Department of Economics, Copenhagen Business School, Frederiksberg, Denmark
e-mail: drm.eco@cbs.dk

Spherical Separation in Machine Learning

Matteo Avolio, Annabella Astorino, and Antonio Fuduli

We extend the spherical separation approach to clustering problems and to the Multiple Instance Learning (MIL) paradigm, the latter constituting a kind of weak supervised classification, using a multisphere criterion. In both the cases, while the centers of the spheres are heuristically fixed, the corresponding radii are computed by solving a specific optimization problem.

In particular, for the clustering problem, all the centers are fixed in advance as the barycenters of the current clusters (as in the standard K-Means algorithm), while the corresponding radii are computed by solving a finite numbers of transportation problems.

Instead, in the case of Multiple Instance Learning (MIL) problem whose objective is to categorize bags of instances, our proposed technique is based on iteratively separating the bags by means of successive maximum-margin spheres (whose number is automatically determined), obtained by solving successive linear programs.

Numerical results on some test problems drawn from the literature show the effectiveness of our proposals.

Keywords: machine learning, spherical separation, clustering, multiple instance learning

Matteo Avolio
Department of Mathematics and Computer Science, University of Calabria, Rende, Italy
e-mail: matteo.avolio@unical.it

Annabella Astorino
ICAR - National Research Council, Rende, Italy, e-mail: annabella.astorino@icar.ncr.it

Antonio Fuduli
Department of Mathematics and Computer Science, University of Calabria, Rende, Italy
e-mail: antonio.fuduli@unical.it

Model Extraction Based on Counterfactual Explanations

Cecilia Salvatore and Veronica Piccialli

Automated decision-making classification systems based on Machine Learning algorithms are often used in many real-life scenarios such as healthcare, credit, or criminal justice. There is thus increasing interest in making Machine Learning systems trustworthy: interpretability, robustness, and fairness are often essential requirements for the deployment of these systems [1]. In particular, according to the European Union's General Data Protection Regulation (GDPR), automated decision-making systems should guarantee the "right to explanations" [2], meaning that those affected by the decision may require an explanation. Counterfactual Explanations are becoming a de-facto standard for a post-hoc explanation [3]. Given an instance of a classification problem, belonging to a class, its counterfactual explanation corresponds to small perturbations of that instance that allow changing the classification outcome. The objective of this work is to try and exploit the information revealed by a small set of examples with their counterfactual explanations to build a surrogate model of the classification system. The idea is to define an optimization problem that provides in output a Forest of Optimal Trees as close as possible to the original classification model, given the information derived from the counterfactual points. This tool can be used, on one hand, to attack a target model; on the other hand, it is also possible to improve the target system by building a more interpretable and sparse model. Preliminary results show the viability of this approach.

Keywords: optimal classification trees, milp, interpretable machine learning

References

1. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
2. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag. **38**, 50–57 (2017)
3. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)

Cecilia Salvatore
University of Rome Tor Vergata, Department of Civil Engineering and Computer Science,
e-mail: cecilia.salvatore@uniroma2.it

Veronica Piccialli
Sapienza University of Rome, Department of Computer, Control and Management Engineering,
e-mail: veronica.piccialli@uniroma1.it

Isolation Forests for Symbolic Data as a Tool for Outlier Mining

Andrzej Dudek and Marcin Pelka

The intuitive definition of an outliers would be 'an observation which deviate so much from other observations as to arouse suspicions that is generated by a different mechanism'. Detection of outliers is a fundamental issue in data analysis, its main goal is to detect and remove anomalous objects from the data. Because the technology changes rapidly, number of databases and their size grows over time, which makes outlier detection and removal even harder.

The main aim of the paper is to propose an adaptation of well-known isolation forest methods, that has been proofed to be useful in outlier detection in the classical data, for symbolic data case. Isolation forest uses well-known decision tree idea to detect anomalies using isolation on the basis how far a data point is to the rest of the data, rather than modelling normal points. The same ideas (applying ideas of the decision tree) can be used for symbolic data. In the empirical part of the presentation artificial and real data with outliers is used to evaluate proposed approach and compare it with DBSCAN, decision tree and kernel discriminant analysis for symbolic data.

Keywords: outliers, symbolic data analysis, isolation forest

References

1. Diday, E., Noirhomme-Fraiture, M.: Symbolic data analysis and the SODAS software. Wiley, Chichester (2008).
2. Liu, F. T., Ting, K. M., & Zhou, Z. H.: Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE (2008).
3. Aggrawal, Ch. C.: Outlier analysis. Second edition. Springer (2018).

Andrzej Dudek
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: andrzej.dudek@ue.wroc.pl

Marcin Pełka
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: marcin.pelka@ue.wroc.pl

Symbolic Clustering Methods Applied to Interval Estimates of Production Cost Quantiles

Dominique Desbois

The decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation, but also on the performance of the productive systems used by the farms, assessing their environmental impacts through the cost of the specific resources used. This communication presents applications of the symbolic clustering methods, proposed in [1, 2] for interval data, to conditional quantile estimates of production costs in agriculture. The interval data clustering tools are used to obtain typologies of European countries and regions, on the basis of the conditional quantile distributions of agricultural production cost empirical estimates. A procedure to find optimal partitions along with various criteria is used. Some standard statistical tests are proceed. This work extends preliminary results published in [3]. The comparative analysis of the econometric results for the main products between European countries and regions illustrates the relevance of the typologies obtained for national and international comparisons based on their specific input productivity.

Keywords: symbolic clustering, interval data, quantile estimates, agricultural production costs, micro-economics

References

1. Carvalho, F., Souza, R., Chavent, M., Lechevallier, Y.: Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, **27(3)**, 167–179 (2006)
2. Chavent, M., Lechevalier, Y., Briant, O.: DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, **52(2)**, 687–701 (2007)
3. Desbois, D.: Applying interval PCA and clustering to quantile estimates: empirical distributions of fertilizer cost estimates for yearly crops in European Countries. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **7(4)**, 695–716 (2021)

Dominique Desbois
Paris-Saclay Applied Economics, INRAE-AgroParisTech, Universitié Paris-Saclay
e-mail: dominique.desbois@inrae.fr

Fisher Discriminant Analysis for Interval Data

Diogo Pinheiro, Maria do Rosário Oliveira, Igor Kravchenko, and Lina Oliveira

In Data Science, entities are typically characterized by vectors of single-valued measurements, called conventional data. Symbolic Data Analysis can model more complex data structures like lists, intervals, histograms, or even distributions. These complex structures may result from the aggregation of conventional data according to the research interests or may exist in their own right. The complexity of these data structures brings new statistical challenges and the need of new methodologies to extract information from it, of which classification is a good example.

In this work, we propose an extension of the conventional Fisher Discriminant Analysis based on Mallows' distance and Moore's interval algebraic structure. The squared Mallows' distance between two interval-valued vectors is written as an explicit form of the sum of two squared Euclidean distances between the vectors of the interval's centers and the vectors of the interval's ranges. The ranges' contribution is weighted according to the assumption of the micro-data distribution within intervals, extending previous work on this topic. This allows us to define associated symbolic covariances matrices that can be decomposed into *within* and *between* covariance matrices. Using Moore's algebraic structure, we generalize Fisher's objective function to interval data, aiming to discriminate between two classes. Examples based on real problems are used to illustrate the advantages of this approach over the conventional one, which ignores the interval structure of the data.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia, Portugal, through the projects UIDB/04621/2020, PTDC/EGE-ECO/30535/2017 and UID/MAT/04459/2020.

Keywords: symbolic data analysis, classification, symbolic Fisher discriminant analysis, interval-valued data, Mallows' distance

Diogo Pinheiro
Instituto Superior Técnico, Universidade de Lisboa, Portugal,
e-mail: diogo.pinheiro.99@tecnico.ulisboa.pt

Maria do Rosário Oliveira
CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal, e-mail: rosario.oliveira@tecnico.ulisboa.pt

Igor Kravchenko
Instituto Superior Técnico, Universidade de Lisboa, Portugal,
e-mail: igor.kravchenko@tecnico.ulisboa.pt

Lina Oliveira
CAMGSD and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal, e-mail: lina.oliveira@tecnico.ulisboa.pt

Stability of Mixed-type Cluster Partitions for Determination of the Number of Clusters

Rabea Aschenbruck, Gero Szepannek, and Adalbert Wilhelm

For partitioning clustering methods, the number of clusters has to be determined in advance and is therefore an important issue. Besides the application of so-called internal validation indices, the determination of an optimal number of clusters can be achieved with stability indices. In this paper several stability based validation methods are investigated with regard to the *k-prototypes* algorithm for mixed-type data, which is an extension to the popular *k-means* algorithm.

Under consideration are the Jaccard coefficient, the Simple Matching coefficient and the indices published by Fowlkes and Mallows as well as by von Luxburg. Furthermore, the weighted average of cluster-wise stability values based on the Jaccard index and a numerical approximation of the stability value interpretation proposed by Ben-Hur et al. were investigated. The stability based approaches are compared to common internal validation indices in a comprehensive simulation study in order to analyze preferability as a function of the underlying data generating process.

Keywords: cluster stability, cluster validation, mixed-type data

Rabea Aschenbruck
Stralsund University of Applied Sciences, Zur Schwedenschanze 15, 18435 Stralsund
e-mail: rabea.aschenbruck@hochschule-stralsund.de

Gero Szepannek
Stralsund University of Applied Sciences, Zur Schwedenschanze 15, 18435 Stralsund
e-mail: gero.szepannek@hochschule-stralsund.de

Adalbert F.X. Wilhelm
Jacobs University Bremen, Campus Ring 1, 28759 Bremen
e-mail: A.Wilhelm@jacobs-university.de

Multinomial Multilevel Models with Discrete Random Effects: a Multivariate Clustering Tool

Chiara Masci, Francesca Ieva, and Anna Maria Paganoni

We propose a Semi-Parametric Mixed-Effects Multinomial regression model to deal with estimation and inference issues in the case of categorical data with a hierarchical structure [1]. Considering a $K-$categories response, the proposed modelling assumes the probability of each response category to be identified by a set of fixed and random effects parameters, one for each logit, estimated by means of an EM algorithm [2]. Random effects are assumed to follow a discrete distribution with an a priori unknown number of support points, that identifies a latent structure at the highest level of grouping, where observations are clustered into $(K-1)-$dimensional subpopulations. This method is an extension of the MSPEM algorithm proposed in [4], in which we relax the independence assumption across random-effects relative to different response categories. Since the category-specific random effects arise from the same subjects, their independence assumption is seldom verified in real data and, by relaxing it, the proposed method properly fits the natural data structure, as emerges by the results of simulation and case studies. In the case study, we apply the algorithm to Politecnico di Milano data, to model different categories of student careers, where students are enrolled in 20 engineering degree courses. Results are compared to the ones of the parametric MCMCglmm appraoch [3].

Keywords: mixed-effects models, categorical responses, discrete random effects, higher education.

References

1. Agresti, A.: An introduction to categorical data analysis. Wiley (2018)
2. Dempster, A. P., Laird, N. M., & Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B, 39(1), 1-22 (1977).
3. Hadfield, J. D.: MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of statistical software, 33, 1-22 (2010).
4. Masci, C., Ieva, F., & A.M. Paganoni: Semiparametric Multinomial Mixed-Effects Models: A University Students Profiling Tool. Annals of Applied Statistics, in press (2022)

Chiara Masci
Politecnico di Milano, via Bonardi 9, 20133 Milano, e-mail: chiara.masci@polimi.it

Francesca Ieva
Politecnico di Milano, via Bonardi 9, 20133 Milano, e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
Politecnico di Milano, via Bonardi 9, 20133 Milano, e-mail: anna.paganoni@polimi.it

PD-clustering for Mixed Data Type

Francesco Palumbo and Cristina Tortora

Data clustering aims to find homogeneous groups in the data using systematic numerical methods; non-hierarchical algorithms can offer considerable advantages over other approaches. Above all, they are easily parallelizable. In a few words, they solve an optimization problem to find two quantities: the cluster memberships and the cluster parameters, which depend on each other. Therefore, algorithms alternatively compute the two quantities, optimizing the same criterion at each step, and stop when the criterion reaches a minimum (maximum). The membership can be *crisp* or *probabilistic*: a point is assigned to all the clusters with a degree of membership. Probabilistic Distance Clustering (PDC) maximizes the classifiability of all the observations assuming that the belonging probability to each cluster is inversely proportional to the distance from the cluster center [1].
To jointly consider mixed data variables, one possible solution is to re-code all variables in a single data type through pre-processing, which can seriously weaken the true association structure. Some satisfactory clustering methods for mixed data exist, but they tend to be slow. The primary issue in clustering mixed data is the identification of a unified similarity metric. The most popular approaches based on this idea are k-prototypes [3] and KAy-means for MIxed LArge data (Kamila) [2]. This proposal extends the PDC to mixed-type data using a dissimilarity for mixed-type data and redefining the cluster centers. The cluster parameters that optimize the criterion are based on the updated dissimilarity and integrated into the algorithm. The performance of the new algorithm are compared to K-prototype and Kamila.

Keywords: probabilistic distance clustering, mixed data

References

1. A. Ben-Israel, C. Iyigun, Probabilistic d-clustering, J Class 25 (1) (2008) 5–26.
2. A. H. Foss, M. Markatou, kamila: Clustering mixed-type data in R and Hadoop, J Stat Softw 83 (13) (2018) 1–45.
   (2022) 1–11.
3. G. Szepannek, clustmixtype: User-friendly clustering of mixed-type data in r, The R Journal 10 (2) (2018) 200–208.

Francesco Palumbo
Università di Napoli Federico II
Department of Political Sciences, Italy, e-mail: fpalumbo@unina.it

Cristina Tortora
San José State University
Department of Mathematics and Statistics, San José (CA), USA,
e-mail: cristina.tortora@sjsu.edu

Anomaly Detection-based Under-sampling for Imbalanced Classification Problems

You-Jin Park, Chun-Yang Peng, Rong Pan, and Douglas C. Montgomery

In this research, we propose a new anomaly detection-based under-sampling method called ADU to improve the classification performance of imbalanced datasets by effectively removing anomalies, such as outliers and noises. To detect the anomalies in different clusters effectively, three useful approaches are considered. Specifically, to detect the outliers belonging to the majority class, neighborhood-based and density-based outlier detection methods, namely OBN (outlierness based on neighborhood) and DBSCAN (density-based spatial clustering based on noise applications) are used [1, 2]. Finally, to eliminate the borderline noise samples in the majority class (i.e., the majority class samples with low membership probabilities), a membership probability-based under-sampling is proposed with changing the under-sampling rate so that a proportion of majority class samples can be removed.

Keywords: classification, class imbalance, class overlap, under-sampling, membership probability

References

1. Gupta, U., Bhattacharjee, V., Bishnu, P.S.: A New Neighborhood-Based Outlier Detection Technique. In: Nath, V., Mandal, J.K. (eds.) MCCS 2018, pp. 527-534. Springer, Singapore (2018)
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad U. (eds.) KDD 1996, pp. 226-231. AAAI Press, Oregon, USA. (1996)

You-Jin Park
Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan, R.O.C., e-mail: yjpark@mail.ntut.edu.tw

Chun-Yang Peng
AU Optronics, Taichung, Taiwan, R.O.C., e-mail: PatrickCYPeng@auo.com

Rong Pan
School of Computing and Augmented Intelligence, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, Arizona, USA, e-mail: Rong.Pan@asu.edu

Douglas C. Montgomery
School of Computing and Augmented Intelligence, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, Arizona, USA, e-mail: doug.montgomery@asu.edu

Continuous Adaptation to Distribution Drifts Through Continual Learning in Manufacturing

Henrique Siqueira and Onay Urfalioglu

Distribution drifts usually occur in the context of time series processing in manufacturing. The causes for this problem include continual optimization of cutting processes in machining and incremental precision loss from long-term stress exerted on the machine's components. Therefore, it is critical to understand the performance of machine learning (ML) models for anomaly detection in manufacturing under distribution drifts for reliable and robust virtual quality control.

In fact, the generalization capability of ML models is strongly compromised when performing inference under a distribution different from the training data [1]. As a result, these models can have a significant drop in performance with a negative impact on overall equipment effectiveness (OEE). OEE is the gold standard in manufacturing that defines productivity based on availability, performance and quality. In this regard, OEE decreases when workpieces produced under the new distribution are wrongly categorized as faulty workpieces, slowing down the manufacturing process by unnecessary and excessive manual quality measurements.

To avoid this kind of problem in production, anomaly detection systems must be constantly evaluated, if necessary, retrained to the new data distribution and redeployed. Instead of adopting a manual and time-demanding workflow, continual learning approaches could acquire novel information from a continuous data stream. This property can possibly endow those systems with continuous adaptation to contextual distribution drifts [2]. In the present study, we aim to investigate how continual learning concepts can be exploited for the development of anomaly detection systems with continuous adaptation to distribution drifts on real-world manufacturing data.

Keywords: continual learning, anomaly detection, manufacturing

References

1. Liang, W. and Zou, J.: MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In ICLR (2022).
2. Mundt, M., Lang, S., Delfosse, Q. and Kersting, K.: CLEVA-Compass: A Continual Learning EValuation Assessment Compass to Promote Research Transparency and Comparability. In ICLR (2022).

Henrique Siqueira · Onay Urfalioglu
Big Data in Manufacturing, Germany,
e-mail: {henrique.siqueira,onay.urfalioglu}@bigdatainmanufacturing.com

Detecting Anomalies with TADGAN: a Case Study

Inês Oliveira e Silva, Carlos Soares, Arlete Rodrigues, and Pedro Bastardo

Generative Adversarial Networks (GAN) is neural network architecture to generate realistic artificial [1]. The generated data is typically used to learn more accurate models. The approach underlying GAN has been adapted for other tasks. One example are TADGAN, which is an adaptation of GAN to detect anomalies in time series data [2]. A time series anomaly is defined as a timepoint or period where a phenomenon displays an unusual behavior. The TADGAN algorithm can be summarized has: 1. generate a time series that is similar to the original one; 2. periods where the distance between the original and the generated time series is very large are classified as anomalies. The classification is affected by a sensitivity hyperparameter, which is a threshold that defines the minimal distance between the time series for the period to be considered anomalous. In this paper, we empirically evaluate TADGAN on the problem of detecting anomalies in data from sensors of a fire detection system. Preliminary results indicate that the performance of the method depends on the values of the sensitivity hyperparameter.

Keywords: anomaly detection, artificial data, GAN

References

1. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.
2. Alexander Geiger, Dongyu Liu, Sarah Alnegheimish, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Tadgan: Time series anomaly detection using generative adversarial networks. *2020 IEEE International Conference on Big Data (Big Data)*, pages 33–43, 2020.

Inês Oliveira e Silva
Faculdade de Engenharia, Universidade do Porto, Portugal, e-mail: up201806385@edu.fe.up.pt

Carlos Soares
LIACC/Faculdade de Engenharia, Universidade do Porto, Portugal,
e-mail: csoares@fe.up.pt

Arlete Rodrigues
Bosch Portugal, Portugal, e-mail: arlete.rodrigues@pt.bosch.com

Pedro Bastardo
Bosch Portugal, Portugal, e-mail: pedro.bastardo@pt.bosch.com

A Trivariate Geometric Classification of Decision Boundaries for Mixtures of Regressions

Filippo Antonazzo and Salvatore Ingrassia

Mixtures of regressions play a prominent role in regression analysis when it is known the population of interest is divided into homogeneous and disjoint groups. This typically consists in partitioning the observational space into several regions through particular hypersurfaces called decision boundaries. A geometrical analysis of these surfaces allows to highlight properties of the used classifier. In particular, a geometrical classification of decision boundaries for the three most used mixtures of regressions (with fixed covariates, with concomitant variables and random covariates) was provided in case of one and two covariates, under Gaussian assumptions and in presence of only one real response variable. This work aims to extend these results to a more complex setting where three independent variables are considered.

Keywords: mixtures of regressions, decision boundaries, hyperquadrics, model-based clustering

References

1. DeSarbo, W. S., Cron, W. L.: A maximum likelihood methodology for clusterwise linear regression. J. Classif. **5**, 249–282 (1988)
2. Grun, B., Leisch, F.: FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. **28**, 1–35 (2008)
3. Hennig, C.: Identifiablity of models for clusterwise linear regression. J. Classif. **17**, 273–296 (2000)
4. Ingrassia S., Minotti S.C., Vittadini G.: Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. J. Classif. **29**, 363-401 (2012)
5. Ingrassia, S., Punzo, A.: Decision boundaries for mixtures of regressions. J. Korean Stat. Soc. **45**, 295-306 (2016)
6. Wedel, M.: Concomitant variables in finite mixture models. Stat. Neerl. **56**, 362–375 (2002)

Filippo Antonazzo
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France e-mail: filippo.antonazzo@inria.fr

Salvatore Ingrassia
Dipartimento di Economia e Impresa, Università di Catania, Corso Italia 55, 95129 Catania, Italy, e-mail: salvatore.ingrassia@unict.it

Clustering Rainfall by Simulated Annealing for Histogram Symbolic Data

Alejandro Chacón and Javier Trejos

We present the use of simulated annealing for clustering histogram symbolic data [2], by the minimization of a criterion based on a Huygens-type decomposition of inertia defined by Wasserstein distance. An efficient cooling scheme for simulated annealing was implemented [1], with variable length Markov chains, allowing a large exploration in the search space. A simplification of inertia change was found in order to efficiently use Metropolis rule. The algorithm was tested on maximum daily rainfall data sets for last 40 years in Costa Rica, in 14 meteorological stations in the Reventazón river basin. Results were compared to a k-means algorithm [3], with a general improvement in quality.

Keywords: maximum daily rainfall, clustering, simulated annealing, symbolic data, Wasserstein distance

References

1. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines. Wiley, New York (1990)
2. Billard, L., Diday, E.: Clustering Methodology for Symbolic Data. Wiley, New York (2020)
3. Irpino, A., Verde, R., De Carvalho, F.A.T.: Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Appl. **41**, 3351–3366 (2014)

Alejandro Chacón
DLZ Corporation, San José, Costa Rica, e-mail: alechaconv@gmail.com

Javier Trejos
CIMPA & School of Mathematics, University of Costa Rica, San José, Costa Rica
e-mail: javier.trejos@ucr.ac.cr

Statistical Assessment of Youth Inclusion in the National Labour Markets

Beata Bal-Domańska

The situation of youth in the labour market has been supported by national and regional policies for many years. Despite the implemented activities aimed at enhancing their professional standing, in many regions their situation is still unfavourable comparing that of adults [1]. An attempt was taken up to develop the typology of national labour markets regarding the inclusion of youth in terms of dynamics and statistics. The inclusiveness of labour markets was defined as a characteristic of the economy in which access to jobs is similar in all groups of the economically active people. Classification methods (Ward's method) and regression methods determining long-term inclusiveness were used to develop the procedure for the European labour markets typology in terms of youth inclusion [2]. As a result, the countries were divided into 4 groups (favourable labour market for both youth and adults groups of workers; unfavourable for youth; difficult labour market; difficult labour market, particularly for youth) along with an assessment of the youth inclusion level. Based on the conducted analysis, it was concluded that in 2020 as many as 12 countries out of 28 EU countries were included in the difficult labour market, particularly for youth group, another 11 represented the countries whose labour market situation can be characterized as unfavourable for youth, two countries (Latvia, Greece) were classified in the difficult labour market group, i.e. the countries where high unemployment rate was accompanied by high youth inclusion level. Only the following 3 countries were included in the favoured labour market group: Germany, Austria and Denmark.

Keywords: typology, panel data models, regression, labour market, youth, european countries

References

1. Asteriou D., Hall G.Stephen: Applied Econometrics. Macmillan Education, Palgrave, (2016)
2. Choudhry, M. T., Marelli, E., Signorelli, M.: Youth and total unemployment rate: The impact of policies and institutions.Rivista Internazionale Di Scienze Sociali **121**, 63–86 (2013)

Beata Bal-Domańska
Wroclaw University of Economics and Business, Wrocław, Komandorska 118/120
e-mail: beata.bal-domanska@ue.wroc.pl

Barriers to Industry Digitization in Poland from the Perspective of High and Medium-high Technology Sector Enterprises

Elżbieta Sobczak, Marcin Pelka, and Karolina Pokorska

Nowadays digital transformation remains one of the most important trends in the development of the economy. It consists in the absorption of modern technologies, which include, i.a. artificial intelligence, machine learning, 3D printing, cloud computing. Digital transformation is the basis for constructing technological and competitive advantage of an enterprise. Enterprises are affected by many of factors impeding the implementation of digital technologies. Due to diverse operating conditions, individual barriers to digitization may have different significance for various enterprises. The research aims to identify the key barriers to investing in digitization in Poland and relating them to the attributes of enterprises. The research material was collected using the diagnostic survey method and applying the survey technique. The time scope of the research covered 2020. The methods of multivariate data analysis were used, with particular emphasis on correspondence analysis and logit regression.

Keywords: digitalization, barriers of digitalization, enterprises from high and medium-high technology

References

1. Gobble, M. M.: Digitalization, digitalization, and innovation. Research-Technology Managment. 61(4), 56-59 (2018).
2. Khan, S., Khan, S., & Aftab, M.: Digitalization and its impact on economy. International Journal of Digital Library Services. 5(2), 138-149 (2015).
3. Mugge, P., Abbu, H., Michaelis, T., Kwiatkowski, A., & Gudergan G.: Patterns of digitalization. A practical guide to digital transformation. Research-Technology Management. 63(2), 27-35 (2020).

Elżbieta Sobczak
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: elzbieta.sobczak@ue.wroc.pl

Marcin Pelka
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: marcin.pelka@ue.wroc.pl

Karolina Pokorska
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: kjgrzyb@gmail.com

Kernel Smoothing-based Probability Contours for Tumour Segmentation

Wenhui Zhang and Surajit Ray

Statistical imaging together with other machine learning techniques are the epitome of digitalizing healthcare and are culminating towards developing innovative tools for automatic analysis of three-dimensional radiological images — PET (Positron Emission Tomography) images [1]. However, the three major challenges in radiology are: (1) increasing demand for medical imaging (2) decreasing turnaround times caused by mass data (3) diagnostic accuracy that leads to a quantification of images. To address these challenges along with ethical issues regarding the use of Artificial Intelligence in patient care, there is a need to develop a new framework of statistical analysis which can be readily used by clinicians and can be trained with a relatively lower number of samples. Most existing algorithms segments a 2D slice by assigning the grid of pixels into the tumour or non-tumour class. Instead of a pixel-level analysis, we will assume that the true intensity comes from a smooth underlying spatial process which can be modelled by a kernel estimates [2]. In this project, we have developed a kernel smoothing-based probability contour method on PET image segmentation, which provides a surface over images that produces contour-based results rather than pixel-wise results, thus mimicking human observers' behaviour. In addition, our methodology provides the tools for developing a probabilistic approach with uncertainty measurement along with the segmentation. Our method is computational efficient and can produce reproductive and robust results for tumour detection, delineation and radiotherapy planning, together with other complementary modalities, such as CT (Computed tomography) images.

Keywords: medical image segmentation, positron emission tomography, kernel, 3d contouring, multi-modal segmentation

References

1. Hatt, M., et al.: Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Meed. Phys. **44**, e1–e42 (2017)
2. Matioli, L.C., et al.: A new algorithm for clustering based on kernel density estimation. J. Appl. Stat. **45**, 347–366 (2018)

Wenhui Zhang, School of Mathematics and Statistics, University of Glasgow,
e-mail: w.zhang.2@research.gla.ac.uk

Surajit Ray, School of Mathematics and Statistics, University of Glasgow,
e-mail: surajit.ray@glasgow.ac.uk

Parameter Estimation for Mixtures of Linear Mixed Models: the EM, CEM and SEM Algorithms

Luísa Novais and Susana Faria

Parameter estimation for mixture models is one of the main and most complex problems regarding these type of models. In particular, maximum likelihood estimation, via the Expectation-Maximization (EM) algorithm, is among the most used methods for estimating the parameters of mixture models. However, the EM algorithm is known to converge slowly and the initialization strategy to be adopted also plays an important role in the estimation of the parameters by this algorithm. In order to overcome these issues, alternative versions of the EM algorithm have been developed over time. In this work, the performance of the EM algorithm is compared to the performance of two modified versions of this algorithm, the Classification Expectation-Maximization (CEM) algorithm and the Stochastic Expectation-Maximization (SEM) algorithm, in the estimation of the parameters for mixtures of linear mixed models through a simulation study.

Keywords: maximum likelihood estimation, mixture models, simulation study

Acknowledgements The research of L. Novais was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference number SFRH/BD/139121/2018.

References

1. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. Met. **39(1)**, 1–38 (1977)
2. Faria, S., Soromenho, G.: Fitting mixtures of linear regressions. J. Stat. Comput. Sim. **80(2)**, 201–225 (2010)
3. Novais, L., Faria, S.: Comparison of the EM, CEM and SEM algorithms in the estimation of finite mixtures of linear mixed models: a simulation study. Comput. Stat. **36(4)**, 2507–2533 (2021)

Luísa Novais
University of Minho, Centre of Molecular and Environmental Biology and Department of Mathematics, Guimarães, Portugal, e-mail: luisa_novais92@hotmail.com

Susana Faria
University of Minho, Centre of Molecular and Environmental Biology and Department of Mathematics, Guimarães, Portugal, e-mail: sfaria@math.uminho.pt

Genomic Prediction Using Machine Learning Methods: Performance Comparison on Synthetic and Empirical Data

Vanda Lourenço, Joseph Ogutu, Rui Rodrigues, and Hans-Peter Piepho

The accurate prediction of genomic breeding values is central to genomic selection (GS) in both plant and animal breeding studies. Genomic prediction (GP) involves the use of thousands of molecular markers spanning the entire genome and therefore requires methods able to efficiently handle high dimensional data. Machine learning (ML) methods, which encompass different groups of supervised and unsupervised learning methods, are becoming widely advocated for and used in GP studies. Although several studies have compared the predictive performances of individual methods, studies comparing the predictive performance of different groups of methods are rare. However, such studies are crucial for identifying (i) groups of methods with superior predictive performance and assessing (ii) the merits and demerits of such groups of methods relative to each other and to the established classical methods. Here, we comparatively evaluate in terms of predictive accuracy and prediction errors (mean squared and mean absolute prediction errors) the genomic predictive performance of several groups of supervised ML methods. Specifically, *regularized* (regularized, adaptive-regularized and group-regularized), *ensemble* (random forests and stochastic gradient boosting), *instance-based* (support vector machine) and *deep-learning* (feed-forward neural network) methods, using one simulated dataset (animal breeding population; three distinct traits) and three empirical maize breeding datasets (same trait; three distinct years). All the methods showed reasonably high predictive performance for most practical selection decisions. However, our results show that the relative predictive performance of the groups of ML methods depends upon both the data and target traits and that for classical regularized methods, increasing model complexity can incur huge computational cost but does not necessarily always substantially improve predictive accuracy. This rules out selection of one benchmark procedure among ML methods for genomic prediction.

Keywords: genomic prediction, predictive accuracy, snps, supervised ml methods

Vanda Lourenço · Rui Rodrigues
Department of Mathematics, CMA, NOVA School of Science and Technology, Caparica, Portugal,
e-mail: vmml@fct.unl.pt;rapr@fct.unl.pt

Joseph Ogutu · Hans-Peter Piepho
Institute of Crop Science, Biostatistics Unit, University of Hohenheim, Stuttgart, Germany
e-mail: jogutu2007@gmail.com;hans-peter.piepho@uni-hohenheim.de

Pooled Mean and Confidence Interval Estimation Combining Different Sets of Summary Statistics

Flora Ferreira, José Soares, Fernanda Sousa, Filipe Magalhães, Isabel Ribeiro, and Pedro Pacheco

Meta-analysis is increasingly used to combine the results of several studies in order to estimate the outcome of interest. When considering a continuous outcome, commonly to estimate pooled mean and confidence interval sample means and standard deviations of primary studies are necessary, which are absent sometimes. Instead, the median along with various measures of spread are reported. Recently, the task of including the studies with the sample size, the median, the range, and/or the quartiles range, in the pooled mean meta-analysis has been achieved by estimating the sample mean and standard deviation of each study [1]. The methods proposed in [1] and in other previous studies (e.g. [2]) to estimate the missing sample mean and standard deviation have been widely used to meta-analyze the means mostly in medical research. Each of these methods assumes that the sets of summary statistics reported in the studies include the median, the sample size, and as measures of spread the minimum and maximum values and/or the first and third quartiles, not covering all possibilities. Some studies report 10th and 90th percentiles (interdecile range) as an additional or unique measure of spread. In this study, we present an approach to estimate the pooled mean and its confidence interval using the estimated means and standard deviations of studies with different sets of summary statistics as an extension of previous works. Simulation studies were used to evaluate the performance of the existing and proposed approaches. Finally, real data in current market research which is of great value to support companies' decision-making including salary estimate and product price estimate were empirically evaluated.

Keywords: confidence interval, meta-analysis, pool data

References

1. McGrath, S., et al.: Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. Stat. Methods Med. Res. **29** (9), 2520–2537 (2020)
2. Luo, D., et al.: : Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range.Stat. Methods Med. Res. **27**(6), 1785–1805 (2018)

Flora Ferreira · José Soares · Pedro Pacheco
BERD – Bridge Engineering Research and Design, Matosinhos, Portugal
e-mail: \{flora.ferreira,jose.soares,pedro.pacheco\}@berd.eu

Fernanda Sousa · Filipe Magalhães · Isabel Ribeiro
Faculty of Engineering (FEUP), University of Porto, Porto, Portugal, e-mail: fcsousa@fe.up.pt

Prediction of Diabetes via Bayesian Network Classifier from Exposure to Environmental Polluting Chemicals Data

Rosy Oh, Hong Kyu Lee, Youngmi Kim Pak, and Man-Suk Oh

Early prediction of diabetes and identification of risk factors for diabetes may prevent or delay diabetes progression. In this study we developed an interactive online application that provides predictive probabilities of normal, prediabetes, diabetes in 4 years using Bayesian network classifier, a machine learning technique. The Bayesian network (BN) was trained using a dataset from Ansung cohort of the Korean Genome and Epidemiological Study (KoGES) in 2008 with follow-up data in 2012. The dataset contained not only traditional risk factors (current diabetes status, Sex, Age, etc.) for future diabetes but also serum biomarkers (AhRL, MIS-ATP, MIS-ROS) for the level of exposure to environmental polluting chemicals (EPC). Based on accuracy and AUC, Tree augmented BN with 11 variables from feature selection was used as our prediction model. The online application implementing our BN prediction system provided a tool that shows customized diabetes prediction for the users but also allows them to simulate the effects of controlling risk factors on future diabetes. The prediction results from our method showed that the EPC biomarkers had dominant and interactive effects on future diabetes and use of the EPC biomerkers together with commonly used risk factors as predictors substantially improved the prediction performance.

Keywords: diabetes mellitus, glucose intolerance, machine learning, bayesian network, environmental pollutants

Rosy Oh
Department of Industrial & Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea, e-mail: rosy.oh5@gmail.com

Hong Kyu Lee
Department of Internal Medicine, Seoul National University College of Medicine, Seoul 110799, Korea,
e-mail: hkleemd@snu.ac.kr

Youngmi Kim Pak
Department of Physiology, Kyung Hee University, College of Medicine, Seoul 02447, Korea
e-mail: ykpak@khu.ac.kr

Man-Suk Oh
dDepartment of Statistics, Ewha Womans University, Seodaemun-Gu, Seoul 03760, Korea,
e-mail: msoh@ewha.ac.kr

Model Performance Metrics for Sample Selection Bias Correction by Pseudo Weighting

An-Chiao Liu, Ton de Waal, Katrijn Van Deun, and Sander Scholtus

Many data sets have not been obtained with a known sampling mechanism and are termed as non-probability samples. Non-probability samples are often treated as a simple random sample when estimating a population parameter (e.g., population mean), and therefore may suffer from sample selection bias. To correct for selection bias, one approach is the pseudo-weighting method [1]. The pseudo-weighting assumes that an inclusion mechanism exists and the inclusion probabilities (propensities) may be estimated. The inverse of the estimated propensities are treated as the pseudo weights.

However, weighting methods in general are sensitive to large variance. The resulting mean squared error after correction may be even larger than the mean squared error before correction [2]. A model assessment framework is needed for selection bias correction by weighting methods. Therefore, it is important to find a suitable indicator of model performance to evaluate the pseudo weights. Some performance metrics that may be useful for evaluation are discuss in this research.

Keywords: sample selection bias, non-probability sample, performance metric

References

1. Elliott, Michael R., and Richard Valliant. Inference for nonprobability samples. Statistical Science 32.2 (2017): 249-264.
2. Meng, Xiao-Li. Statistical Paradises and Paradoxes in Big Data (I) Law of Large Populations, Big data Paradox, and the 2016 US Presidential Election. The Annals of Applied Statistics 12.2 (2018): 685-726.

An-Chiao Liu
Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands,
Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands,
e-mail: a.liu@uvt.nl

Ton de Waal
Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands,
Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands
e-mail: t.dewaal@cbs.nl

Katrijn Van Deun
Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands e-mail: k.vandeun@uvt.nl

Sander Scholtus
Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands,
e-mail: s.scholtus@cbs.nl

Some Factors that Influence the Nature of Road Traffic Accidents

Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Leonor Rego, Vitor Nogueira, Paulo Quaresma, José Saias, Daniel Santos, Pedro Nogueira, Marcelo Silva, Rosalina Pisco Costa, Patrícia Gois, and Paulo Rebelo Manuel

Road traffic accidents are one of the major social problems in modern societies. Data science plays an important role in its explanation and prediction. One of the main objectives of accident data analysis is to identify the main factors associated with a road traffic accident. The present study aims to contribute to the identification of the determinants for the nature (collision, car crash or pedestrian runing over) of road accidents. Four-year road accident data from 2016 to 2019 in a district of Portugal is analyzed. Three logit models, a multinomial logit regression model and several machine learning algorithms are used, and their performance is compared. Findings show that some determinants which can explain the nature of the accident are: geographical factors (municipality, in/outside localities and parking areas), temporal factors (air temperature and wheater), time of the day (hour, day of the week and month), drivers' characteristics (gender and age), vehicles' features (type and age) and road characteristics (straight/curved track and road type).

Keywords: generalized linear models, machine learning, road traffic accidents

---

Paulo Infante, Gonçalo Jacinto and Anabela Afonso
CIMA, Depto. de Matemática, Universidade de Évora,
e-mail: pinfante@uevora.pt, gjcj@uevora.pt, aafonso@uevora.pt

Lenor Rego and Daniel Santos
Universidade de Évora, e-mail: lrego@uevora.pt, dfsantos@uevora.pt

Vitor Nogueira and José Saias
Algoritmi Research Centre, Depto. de Informática, Universidade de Évora,
e-mail: vbn@uevora.pt, jsaias@uevora.pt

Paulo Quaresma
Algoritmi Research Centre, Universidade de Évora, e-mail: pq@uevora.pt

Pedro Nogueira and Marcelo Silva
ICT, Depto. de Geociências, Universidade de Évora,
e-mail: pmn@uevora.pt, marcelogs@uevora.pt

Rosalina Costa
CICS.NOVA.UEVORA, Depto. de Sociologia, Universidade de Évora,
e-mail: rosalina@uevora.pt

Patrícia Gois
Depto. de Artes Visuais e Design, Universidade de Évora, e-mail: pafg@uevora.pt

Paulo Rebelo Manuel
CIMA, Universidade de Évora, e-mail: pjsrm@uevora.pt

Comparing Variable Selection Methods for High-dimensional Compositional Data in a Discriminant Analysis Context

Pepus Daunis-i-Estadella, Glòria Mateu-Figueras, Viktorie Nesrstová, Karel Hron, and Josep A. Martín-Fernández

Orthonormal logratio coordinates are used for statistical analysis of compositions. Although there is an infinite number of potential orthonormal bases, using the Principal balances (PB) algorithms a specific basis is constructed for maximizing the retained variance [3]. PB can be used for selecting some parts in order to reduce the data dimension. This is particularly important in the context of high-dimensional compositional data (CoDa), such as omic data (genomic, proteomic or metabolomic) where the identification of biomarkers that allow to discriminate disease has great interest. This work focuses on the comparison of different proposals for variable selection in order to apply a discriminant analysis (DA): the Selbal approach [5], a forward-selection method for the identification of a balance associated with the variable of interest; the CLR-lasso and CoDa-lasso, a penalised regression approach [6]; the approaches proposed in [2] based on the index extracted of the variation matrix and average accuracy using Partial Least Squares (PLS) DA; the stepwise supervised methods for selecting pairwise logratios [1] and the variable selection in CoDa using PLS-PB coordinates, a modified procedure consisting in the application of PLS regression [4]. Simulated data sets and a real data set with the microbiota associated to the colonic mucosa of patients with different subtypes of inflammatory bowel disease are used for illustrating the performance of the methods.

Keywords: variable selection, discriminant analysis, principal balances

References

1. Coenders, G., Greenacre, M.: Three approaches to supervised learning for compositional data with pairwise logratios. arXiv:2111.08953 [stat.ML] (2021)
2. Filzmoser, P., Hron, K., Templ, M.: Applied compositional data analysis. Springer (2018)
3. Martín-Fernández J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R.: Advances in principal balances for compositional data. Mathematical Geosciences **50**, 273–298 (2018)
4. Nesrstová V, Hron, K., Martín-Fernández J.A, Filzmoser, P., Palarea-Albaladejo, J., Wilms, I.: Variable selection in compositional data using PLS-based balance coordinates. Data Science, Statistics & Visualisation (DSSV), 7-9 July (2021)
5. Rivera-Pinto J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L.: Balances: a new perspective for microbiome analysis. mSystems **3**, e00053-18 (2018)
6. Susin, A., Wang, Y., Lê Cao, K.-A., Calle, M. L.: Variable selection in microbiome compositional data analysis. NAR Genomics and Bioinformatics **2**, lqaa029 (2020)

P. Daunis-i-Estadella, G. Mateu-Figueras, J.A. Martín-Fernández
University of Girona, Campus Montilivi, 17003 Girona, Spain; e-mail: gloria.mateu@udg.edu

V. Nesrstová , K. Hron
Palacký University, 17. Isitopadu 12, 77146 Olomouc, the Czech Republic

Cluster Analysis and Genetic Risk Score in Age-related Macular Degeneration - the Coimbra Eye Study

Rita Coimbra

Age-related macular degeneration (AMD) is a complex multifactorial disease strongly influenced by a combination of genetic and environmental factors and it is the leading cause of severe vision loss in people over 55 years in developed countries. To assess the individual AMD risk, a genetic risk score (GRS) weighted by the effect size of 52 genetic variants identified by the IAMDGC GWAS was calculated [1]. Despite the significantly differences in GRS among controls and AMD cases, it was not possible to distinguish them based on GRS only.
In this work we explore pathway-specific GRS for the complement system, ARMS2 gene, lipid metabolism and extracellular (ECM) matrix remodelling in a sample of 824 caucasian individuals from the central region of Portugal (Mira) [2]. To explore whether pathway-based genotype and environmental data can separate AMD subgroups, we used k-medoids clustering algorithm with Gower distance. The silhouette coefficient was used to determine the optimal number of clusters.
Four clusters were identified, with a similar median age and body mass index (BMI) in each cluster. The cluster with the highest GRS value for the complement system pathway was cluster 4, for the ARMS2 and the ECM pathways was cluster 2 and for lipid-based GRS were clusters 2 and 3. Cluster 4 has the highest overall GRS, strengthening the contribution of the complement system to the AMD genetic risk. Most severe AMD stages (intermediate and late AMD cases) were more prevalent in cluster 3. Despite the interesting results and to fully discriminate AMD subgroups, lifestyle and nutritional contribution should be included in the model.

Keywords: age-related macular degeneration, genetic risk score, clustering

References

1. de Breuk, A. *et al*: Development of a Genotype Assay for Age-Related Macular Degeneration: The EYE-RISK Consortium. Ophthalmology **128(11)**, 1604-1617 (2021)
2. Farinha, C.V.L., Cachulo, M.L., Alves, D. *et al*: Incidence of Age-Related Macular Degeneration in the Central Region of Portugal: The Coimbra Eye Study - Report 5. Ophthalmic Res. **61(4)**, 226-235 (2019)

Rita Coimbra
Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal (AIBILI), e-mail: racoimbra@aibili.pt

Sensor System for Standardizing Articulation Patterns According to Korean Phonemes

Seong Tak Woo and Da Hee Oh

A sensor system (consisting of approximately 70 channels on an artificial plate) was established for articulation pattern standardization, and the characteristics of Korean speech were analyzed. Articulation pattern information was obtained using a sensor with a group of healthy participants. The measured signals were analyzed using a bio-signal (tongue contact) processing module and viewer program. Notably, the data were obtained from only two healthy adults, and certain limitations remain in the application of data analysis technologies (such as neural network/regression model). As a preliminary study, we conducted repeated measures analysis of variance to determine the difference in the closure time and contact area according to the utterance unit, meaning, and syllable structure. Specifically, we attempted to determine the presence of the preceding consonant (C) in two successive consonants (VCCV) with the same articulation place. Results of experiments demonstrated that the articulation sensor and analysis method can be applied to standardize pronunciation patterns. Although extensive healthy group data and analytical techniques remain to be identified, the findings can provide insights regarding initial experimental setup (proposed sensor, acquisition method, etc.) for standardization studies.

Keywords: articulation pattern, tongue strength, analysis of variance, speech therapy, communication disorders

References

1. Dagenais, P.A., Lorendo, L.C., McCutcheon, M.J.: A study of voicing and context effects upon consonant linguapalatal contact patterns. Journal of Phonetics **22**, 225–238 (1994)
2. Braislin, M.A.G., Cascella, P.W.: A preliminary investigation of the efficacy of oral motor exercises for children with mild articulation disorders. International Journal of Rehabilitation Research **28**, 262–266 (2005)

Seong Tak Woo
Dong Seoul University, 76, Bokjeong-ro, Sujeong-gu, Seongnam-si, Gyeonggi-do, S. Korea,
e-mail: stwoo@du.ac.kr

Da Hee Oh
Daegu University, 201, Daegudae-ro, Gyeongsan-si, Gyeongsangbuk-do, S. Korea,
e-mail: duhee03@naver.com

Categorical Data Visualization and the Cressie-Read Divergence Statistic

Eric J. Beh and Rosaria Lombardo

The literature on correspondence analysis is largely centred on Pearson's chi-squared statistic (Greenacre, 1984; Beh and Lombardo, 2014). For such an analysis, differences between categories is assessed using the chi-squared distance while distances in a low-dimensional space are Euclidean. Recently, Beh and Lombardo (2022) showed that the Cressie-Read divergence statistic (Cressie and Read, 1984) plays a pivotal role in correspondence analysis. This is because Pearson's statistic, the Freeman-Tukey statistic, the log-likelihood ratio statistic and others are special cases of this statistic. Therefore, differences amongst categories can be assessed using this divergence statistic and still yield Euclidean distances in a low-dimensional space. This is very appealing since it provides a greater range of difference measures, including the chi-squared, Hellinger and logarithmic distances.This paper compares the quality and configuration of points in a low-dimensional correspondence plot using the Cressie-Read divergence statistic. This will be done by deriving a global measure that allows for such a comparison and is an extension of a similar measure presented in the past; see, for example, Cuadras, Cuadras and Greenacre (2006).

Keywords: Pearson's chi-squared, Cressie-Read divergence statistic, comparison

References

1. Beh, E. J., Lombardo, R.: Correspondence analysis: Theory, practice and new strategies. Wiley, Chichester (2014)
2. Beh, E. J., Lombardo, R.: Correspondence analysis and the Cressie-Read Divergence Statistic. Submitted (2022)
3. Cressie, N. A. C., Read, T. R. C.: Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B, **46**, 440–464 (1984)
4. Cuadras, C.M., Cuadras, D., Greenacre, M. J.: A Comparison of Different Methods for Representing Categorical Data. Communications in Statistics - Simulation and Computation, **35**, 447–459 (2006)
5. Greenacre, M.: Theory and application of correspondence analysis. Academic Press, London (1984)

Eric J. Beh
School of Information & Physical Sciences, University of Newcastle, Newcastle, and University of Wollongong (Aust) & Stellenbosch University (SA), e-mail: eric.beh@newcastle.edu.au

Rosaria Lombardo
Department of Economics, University of Campania "L. Vanvitelli",
e-mail: rosaria.lombardo@unicampania.it

Biplots Based on Latent Variable Models in the Analysis of Ecological Communities

Jenni Niku and Sara Taskinen

Throughout the last decade, joint species distribution modelling concept (JSDM, [1]) has gained popularity in the analysis of species communities, where the object of interests is usually in describing and understanding the structure and dynamics of a group of interacting species. JSDMs offer a flexible approach for determining and assessing the mechanisms that drive the communities to exhibit patterns in observed species communities. One particular JSDM approach, which we focus on, is based on generalized linear latent variable models (GLLVMs,[2]). The GLLVMs are closely related to latent factor models which are applicable in several fields of science, but here we consider them in the analysis of species communities.

While being capable to model, for example, environmental factors on species abundances and species-to-species associations, the GLLVMs also provide a model-based tool for producing ordinations and biplots, thus providing a powerful tool for visualizing those species-to-species associations and connecting them to the environmental or underlying latent factors driving the species communities. We introduce the GLLVMs and show how they can be used for producing biplots. In addition, some useful properties of methods are illustrated by applying them to an ecological dataset.

Keywords: biplot, latent variable model, joint species distribution model, species community

References

1. Warton, D.I., Guillaume Blanchet, F., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K.C.: So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution. **30**(12), 766–779 (2015)
2. Niku, J., Warton, D.I., Hui, F.K.C., Taskinen, S.: Generalized linear latent variable models for multivariate count and biomass data in ecology. Journal of Agricultural, Biological and Environmental Statistics, **22**(4), 498–522 (2017)

Jenni Niku
Department of Biological and Environmental Science, University of Jyvaskyla, Finland
e-mail: jenni.m.e.niku@jyu.fi

Sara Taskinen
Department of Mathematics and Statistics, University of Jyvaskyla, Finland
e-mail: sara.taskinen@jyu.fi

Biplot Representation of Partial Least Squares Regression for Binary Responses

Laura Vicente-Gonzalez and Jose Luis Vicente-Villardon

In this work we describes biplot representation for visualization of Partial Least Squares Regression for Binary Responses (PLS-BLR). The PLS-BLR is a generalization of Partial Least Squares Regression [3] to handle a matrix of continuous predictors and a set of binary responses. The resulting biplot will be a combination of a traditional representation for numeric data and a Logistic Biplot as described by [2] or [1].

First we describe the base methods, PLSR, PLSR-BLR and their associated biplots, then the connection among them. Finally we present an application to real data.

Software packages for the calculation of the main results are also provided.

Keywords: biplot, binary data, PLS, PLS-BLR

References

1. Demey, J., Vicente-Villardon, J. L., Galindo, M. P., Zambrano, A.: Identifying Molecular Markers Associated With Classification Of Genotypes Using External Logistic Biplots. Bioinform., **24(24)**,2832-2838, 2008
2. Vicente-Villardon, J.L., Galindo, M.P., Blazquez-Zaballos, A.: Logistic biplots. In: Greenacre, M., Blasius, J. (eds.) Multiple Correspondence Analysis and related methods., pp. 503-521. Chapman and Hall, New York (2006)
3. Wold., H.: Soft modeling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. J. Appl. Probab. , **12(S1)**, 117-142, (1975)

Laura Vicente-Gonzalez
Department of Statistics, University of Salamanca, C/Alfonso X el Sabio S/N, 37007 Salamanca, Spain, e-mail: laura20vg@usal.es

Jose Luis Vicente-Villardon
Department of Statistics, University of Salamanca, C/Alfonso X el Sabio S/N, 37007 Salamanca, Spain, e-mail: villardon@usal.es

Generalized Spatio-temporal Regression with PDE Penalization

Eleonora Arnone, Elia Cunial, and Laura M. Sangalli

We develop a novel generalised linear model for the analysis of data distributed over space and time. The model involves a nonparametric term f, a smooth function over space and time. The estimation is carried out by the minimization of an appropriate penalized negative log-likelihood functional, with a roughness penalty on f that involves space and time differential operators, in a separable fashion, or an evolution partial differential equation. The model can include covariate information in a semi-parametric setting. The functional is discretized by means of finite elements in space, and B-splines or finite differences in time. Thanks to the use of finite elements, the proposed method is able to efficiently model data sampled over irregularly shaped spatial domains, with complicated boundaries. To illustrate the proposed model we present an application to study the criminality in the city of Portland, from 2015 to 2020.

Keywords: functional data analysis, spatial data analysis, semiparametric regression with roughness penalty

Eleonora Arnone
Dipartimento di Scienze Statistiche, Università di Padova, Via Cesare Battisti, 241, 35121 Padova, Italy, e-mail: eleonora.arnone@unipd.it

Elia Cunial
Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: elia.cunial@mail.polimi.it

Laura M. Sangalli
Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: laura.sangalli@polimi.it

Impact Point Selection in Semiparametric Bifunctional Models

Silvia Novo, Germán Aneiros, and Philippe Vieu

Nowadays, most applied sciences have to deal with datasets containing one, or more, functional object. Thus, developing techniques for functional data analysis with high level of flexibility and interpretability has become a target in statistical research.

Accordingly, a new sparse semiparametric model is proposed, which incorporates the influence of two functional random variables in a scalar response in a flexible and interpretable manner. One of the functional covariates is included through a single-index structure, while the other is included linearly through the high-dimensional vector formed by its discretised observations. Due to the sparse nature of the linear component, variable selection is needed (see [1] for a review). The problem is that standard variable selection methods (such that the proposed in [2]) can provide inadequate results. Then, two new algorithms are presented for selecting impact points in the linear component and estimating the model (see [3] for details). Both procedures utilise the functional origin of linear covariates. Finite sample experiments demonstrated the scope of application of both algorithms. Some asymptotic results support both procedures. A real data application showed the applicability of the presented methodology from a predictive perspective and low computational cost.

Keywords: functional data analyisis, variable selection, semiparametric regression

References

1. Aneiros, G., Novo, S., Vieu, P.: Variable selection in functional regression models: A review, J. Multivariate Anal., **188**, 104871, (2021)
2. Novo, S., Aneiros, G., Vieu, P.: Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. Test, **30**, 481–504 (2021)
3. Novo, S., Vieu, P., Aneiros, G.: Fast and efficient algorithms for sparse semiparametric bifunctional regression, Aust. N. Z. J. Stat., **63**, 606–638 (2021)

Silvia Novo
Research group MODES, Department of Mathematics, CITIC, Universidade da Coruña, A Coruña, Spain, e-mail: s.novo@udc.es

Germán Aneiros
Research group MODES, Department of Mathematics, CITIC, Universidade da Coruña, A Coruña, Spain, e-mail: german.aneiros@udc.es

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier–Toulouse III, Toulouse, France
e-mail: philippe.vieu@math.univ-toulouse.fr

Latent Function-on-scalar Regression Models for Observed Sequences of Correlated Binary Data: a Restricted Likelihood Approach

Fatemeh Asgari and Valeria Vitelli

In function-on-scalar regression problems, the response curve is sometimes observed as a sequence of correlated binary or multilevel data. This kind of situations can be handled via the family of generalized functional regression models, with several proposals in this direction already present in the literature [2, 3]. In this talk, we introduce a functional regression setting where the random response curve is unobserved, and only its dichotomized version as a sequence of correlated binary data is observed. We propose a practical computational framework for maximum likelihood analysis relying on the use of a complete data likelihood, which has the advantages of scaling to large datasets, and of handling non-equally spaced and missing observations effectively and flexibly. The proposed method is used in the Function-on-Scalar regression setting, with the latent response variable being a Gaussian random element taking values in a separable Hilbert space. We provide smooth estimations for the functional regression coefficients and principal components by introducing an adaptive Monte Carlo Expectation Maximization (MCEM) algorithm that circumvents selecting the smoothing parameters. The novel method is described in details in [1], where its performance is also demonstrated by simulated and real case studies, and is implemented in the `R` package `dfrr`.

Keywords: function-on-scalar regression, longitudinal binary data, mcem algorithm

References

1. Asgari, F., Alamatsaz, M.H., Vitelli, V., Hayati, S.: Latent function-on-scalar regression models for observed sequences of binary data: a restricted likelihood approach. arXiv preprint arXiv:2012.02635.(2020)
2. Goldsmith, J., Zipunnikov, V., Schrack, J.: Generalized multilevel function-on-scalar regression and principal component analysis. Biometrics, **71**, 344–353 (2015)
3. Scheipl, F., Gertheiss, J., Greven, S.: Generalized functional additive mixed models. Electron. J. Stat. **10**, 1455–1492 (2016)

Fatemeh Asgari
Institute of Basic Medical Sciences, University of Oslo, Domus Medica, Sognsvannsveien 9, 0372, Oslo, e-mail: `fatemeh.asgari@medisin.uio.no`

Valeria Vitelli
Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Domus Medica, Sognsvannsveien 9, 0372, Oslo, e-mail: `valeria.vitelli@medisin.uio.no`

pcTVI: Parallel MDP Solver Using a Decomposition into Independent Chains

Jaël Champagne Gareau, Éric Beaudry, and Vladimir Makarenkov

Markov Decision Processes (MDPs) are useful to solve real-world probabilistic planning problems [1]. However, finding an optimal solution in an MDP can take an unreasonable amount of time when the number of states in the MDP is large. In this paper, we present a way to decompose an MDP into Strongly Connected Components (SCCs) and to find dependency chains for these SCCs. We then propose a variant of the Topological Value Iteration (TVI) algorithm [2], called *parallel chained TVI* (pcTVI), which is able to solve independent chains of SCCs in parallel leveraging modern multicore computer architectures. The performance of our algorithm was measured by comparing it to the baseline TVI algorithm on a new probabilistic planning domain introduced in this study. Our pcTVI algorithm led to a speedup factor of 20, compared to traditional TVI (on a computer having 32 cores).

Keywords: markov decision process (mdp), planning, strongly connected components, dependancy chains, parallel computing

References

1. Mausam, Kolobov: Planning with Markov Decision Processes: An AI Perspective. Morgan & Claypool (2012)
2. Dai, Mausam, Weld, Goldsmith: Topological value iteration algorithms. J. Artif. Intell. Res., vol. 42, pp. 181–209 (2011)

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

Classification of Viral Pneumonia Images via Multiple Instance Learning

Antonio Fuduli, Matteo Avolio, Eugenio Vocaturo, and Ester Zumpano

We present an application of the Multiple Instance Learning (MIl) paradigm to the classification of pneumonia X-ray images, considering three different categories: radiographies of healthy people, of people with bacterial pneumonia and of people with viral pneumonia. The proposed algorithms, which are very fast in practice, appear promising especially if we take into account that no preprocessing technique on the images has been used.

In particular we will focus on the application of three different MIL instance-space approaches: MIL-RL [1], based on a Lagrangian relaxation technique, mi-SPSVM [2], which combines the classical Support Vector Machine approach with the Proximal Support Vector Machine technique and MIL-kink [3], which provides a separation hyperplane fixing in advance the normal and computing the bias by nonsmooth techniques.

Numerical results on some real-world data sets are presented.

Keywords: multiple instance learning, classification problems, image recognition

References

1. Astorino, A., Fuduli, A., Gaudioso, M.: A Lagrangian relaxation approach for binary multiple instance classification. IEEE Trans. Neural Netw. Learn. Syst. **30**, 2662–2671 (2019)
2. Avolio, M., Fuduli, A.: A semiproximal support vector machine approach for binary multiple instance learning. IEEE Trans. Neural Netw. Learn. Syst. **32**, 3566–3577 (2021)
3. Fuduli, A., Gaudioso, M., Khalaf, W., Vocaturo, E.,: A heuristic approach for multiple instance learning by linear separation. Soft Comput. **26**, 3361–3368 (2022)

Antonio Fuduli
Department of Mathematics and Computer Science, University of Calabria, Rende, Italy,
e-mail: antonio.fuduli@unical.it

Matteo Avolio
Department of Mathematics and Computer Science, University of Calabria, Rende, Italy,
e-mail: matteo.avolio@unical.it

Eugenio Vocaturo
DIMES, University of Calabria, Rende, & CNR NanoTec, Cosenza, Italy,
e-mail: e.vocaturo@dimes.unical.it

Ester Zumpano
DIMES, University of Calabria, Rende, Italy, e-mail: e.zumpano@dimes.unical.it

Nonlinear Approaches for Multiple Instance Learning

Annabella Astorino, Matteo Avolio, and Antonio Fuduli

Multiple Instance Learning (MIL) is a variant of traditional supervised learning consisting in classifying bags of instances. Differently from the traditional supervised learning scenario, each example is not represented by a fixed-length vector of features but by a bag of feature vectors called instances. In the training phase the classification labels are only provided for each entire bag whereas the labels of the instances inside them are unknown. The final task is to learn a model that predicts the labels of the new incoming bags together with the labels of the instances inside them.

We address the MIL problem in the case of two types of instances and two types of bags (positive and negative) through polyhedral approaches. The idea is to generate a polyhedral separation surface such that, for each positive bag, at least one of its instances is inside the polyhedron and all the instances of each negative bag are outside. We come out with two models. For solving the first one, starting from the MIL-SVM type model proposed in [1], we develope a technique based on iteratively separating the bags by means of successive maximum-margin polyhedral surfaces, obtained by solving successive linear programs. In the second, substituting the separating hyperplane with a maximum-margin polyhedral surface in the SVM-type model presented in [2], we obtain a nonsmooth unconstrained optimization problem of DC (Difference of Convex) type that we solve by adapting the DCA algorithm. Numerical results are presented on a set of benchmark datasets.

Keywords: multiple instance learning, SVM, polyhedral separation

References

1. Andrews, S.,Tsochantaridis,I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Becker, S., Thrun, S., Obermayer, K., (eds.) Advances in Neural Information Processing Systems, pp. 561-568. MIT Press, Cambridge (2003).
2. Bergeron, C., Moore, G., Zaretzki, J., Breneman, C.M., Bennett, K.P.: Fast bundle algorithm for multiple-instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1068–1079 (2012).

Annabella Astorino
ICAR, Consiglio Nazionale delle Ricerche, Italy, e-mail: annabella.astorino@icar.cnr.it

Matteo Avolio
Dip. di Matematica e Informatica, Universitá della Calabria, Italy,
e-mail: matteo.avolio@unical.it

Antonio Fuduli
Dip. di Matematica e Informatica, Universitá della Calabria, Italy,
e-mail: antonio.fuduli@unical.it

An Online Minorization-Maximization Algorithm

Hien Duy Nguyen, Florence Forbes, Gersende Fort, and Olivier Cappé

Modern statistical and machine learning settings often involve high data volume and data streaming, which require the development of online estimation algorithms. The online Expectation–Maximization (EM) algorithm extends the popular EM algorithm to this setting, via a stochastic approximation approach. We show that an online version of the Minorization–Maximization (MM) algorithm, which includes the online EM algorithm as a special case, can also be constructed in a similar manner. We demonstrate our approach via an application to the logistic regression problem and compare it to existing methods.

Keywords: expectation–maximization, minorization–maximization, parameter estimation, online algorithms, stochastic approximation

Hien Duy Nguyen
School of Mathematics and Physics, University of Queensland, St. Lucia, 4067 QLD, Australia,
e-mail: h.nguyen7@uq.edu.au

Florence Forbes
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000, Grenoble, France
e-mail: florence.forbes@inria.fr

Gersende Fort
Institut de Mathématiques de Toulouse, CNRS, Toulouse, France
e-mail: gersende.fort@math.univ-toulouse.fr,

Olivier Cappé
ENS Paris, Universite PSL, CNRS, INRIA, France, e-mail: Olivier.Cappe@cnrs.fr

Frugal Gaussian Clustering of Huge Ombalanced Datasets Through a Bin-marginal Approach

Filippo Antonazzo, Christophe Biernacki, and Christine Keribin

Clustering conceptually reveals all its interest when the dataset size considerably increases since there is the opportunity to discover tiny but possibly high value clusters which were out of reach with more modest sample sizes. However, clustering is practically faced to computer limits with such high data volume, since possibly requiring extremely high memory and computation resources. In addition, the classical subsampling strategy, often adopted to overcome these limitations, is expected to heavily failed for discovering clusters in the highly imbalanced cluster case. Our proposal first consists in drastically compressing the data volume by just preserving its bin-marginal values, thus discarding the bin-cross ones. Despite this extreme information loss, we then prove identifiability property for the diagonal mixture model and also introduce a specific EM-like algorithm associated to a composite likelihood approach. This latter is extremely more frugal than a regular but unfeasible EM algorithm expected to be used on our bin-marginal data, while preserving all consistency properties. Finally, numerical experiments highlight that this proposed method outperforms subsampling both in controlled simulations and in various real applications where imbalanced clusters may typically appear, such as image segmentation, hazardous asteroids recognition and fraud detection.

Keywords: imbalanced clustering, large size data, gaussian mixture models, binned data, random subsampling, frugal learning

Filippo Antonazzo
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, e-mail: filippo.antonazzo@inria.fr

Christophe Biernacki
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, e-mail: christophe.biernacki@inria.fr

Christine Keribin
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France, e-mail: christine.keribin@universite-paris-saclay.fr

Reinforced EM Algorithm Through Clever Initialization for Clustering with Gaussian Mixture Models

Joshua Tobin, Chin Pang Ho, and Mimi Zhang

Gaussian mixture models (GMMs) are a prominent clustering method that assume the data generating process to be a mixture distribution of a finite number of Gaussian components. The clusters are taken to be the constituent components. GMMs are ubiquitous in clustering applications as they are both simple and flexible, allowing the clusters to vary in terms of their shape, size and orientation. In practice, the Expectation Maximization (EM) algorithm is used to find maximum likelihood estimates of the GMM parameters. As the likelihood function is non-convex, care must be taken to ensure that EM is initialized with values close to the true parameters. Present initialization methods fail to provide such estimates. The random initialization approach fails to ensure consistency between runs, and can cause EM to converge to arbitrarily bad values of the likelihood. A widely used deterministic approach initializes EM using partitions from likelihood-based hierarchical clustering. This method is computationally infeasible for large datasets, and is ill-suited for detecting clusters of different sizes. We here propose initialization scheme which is applicable to large datasets and reliably produces consistent clustering outputs. We apply an efficient mode-finding criterion to generate a set of initial mean vectors. This set is then pruned through optimization of a convex objective with an adaptive cardinality penalty. We demonstrate how to prune the mean vectors one at a time, generating a sequence of nested clustering results. We provide guidance on how to select the optimal clustering from this sequence. We present theoretical guarantees for the quality of our initialization and experimental results to verify that our algorithm works well in practice.

Keywords: exemplar, Gaussian mixtures, density peaks, convex optimization.

Joshua Tobin
School of Computer Science & Statistics, Trinity College Dublin, Ireland, e-mail: tobinjo@tcd.ie

Chin Pang Ho
School of Data Science, City University of Hong Kong, Hong Kong,
e-mail: clint.ho@cityu.edu.hk

Mimi Zhang
School of Computer Science & Statistics, Trinity College Dublin, Ireland.
I-Form Advanced Manufacturing Research Centre, Science Foundation Ireland, Ireland,
e-mail: mimi.zhang@tcd.ie

Exact Computation of the Angular Halfspace Depth

Stanislav Nagy and Rainer Dyckerhoff

Directional data arise naturally as observations lying in the unit sphere of a d-dimensional Euclidean space. The angular halfspace depth [1] is a nonparametric tool for the analysis of directional data, with wide applications in e.g. classification and pattern recognition tasks. The angular halfspace depth was proposed already in 1987 in [2], but its widespread use has been hampered in practice by significant computational issues. We address these problems by considering a simple projection scheme that allows reducing the computation of the angular halfspace depth to the task of evaluating a variant of the usual halfspace depth in a linear space [3]. Efficient algorithms for exact computation and approximation of the angular halfspace depth are thus developed.

Keywords: angular halfspace depth, computation, directional data analysis

References

1. Liu, R.Y., Singh, K.: Ordering directional data: concepts of data depth on circles and spheres. Ann. Statist. **20**(3), 1468–1484 (1992)
2. Small, C.G.: Measures of centrality for multivariate and directional distributions. Canad. J. Statist. **15**(1), 31–39 (1987)
3. Dyckerhoff, R., Mozharovskyi, P.: Exact computation of the halfspace depth. Comput. Statist. Data Anal. **98**, 19–30 (2016)

Stanislav Nagy
Charles University, Faculty of Mathematics and Physics, Prague, Czech Rep.,
e-mail: nagy@karlin.mff.cuni.cz

Rainer Dyckerhoff
University of Cologne, Institute of Econometrics and Statistics, Köln, Germany,
e-mail: rainer.dyckerhoff@statistik.uni-koeln.de

Reconstruction of Atomic Measure Based on its Simplicial Depth

Petra Laketa and Stanislav Nagy

Statistical depth functions have been introduced in order to generalize the notions of ranks, orderings and quantiles in the multivariate case. Depth is a function that to any point x assigns the quantity which aims to describe how centrally positioned is x respect to a given measure. It is of importance to explore whether the depth function contains all the information of the underlying measure, i.e. does it characterise it. By characterising we mean that there is no other measure with the same depth function everywhere. We focus on the special case of simplicial depth, introduced in [2, 2] and the class of atomic measures. In this particular setup, under mild assumption of atoms being in general position, we prove that the characterisation property is satisfied and describe an algorithm for recovering atomic measure from its simplicial depth.

Keywords: simplicial depth, atomic measures, reconstruction

References

1. Liu , R. Y.: On a notion of simplicial depth. Proc. Natl. Acad. Sci. U.S.A., **85**(6), 1732–1734 (1988)
2. Liu, R. Y.: On a notion of data depth based on random simplices. Ann. Statist., **18**(1), 405-–414 (1990)

Petra Laketa
Charles University, Faculty of Mathematics and Physics, Prague, Czech Rep.
e-mail: laketa@karlin.mff.cuni.cz

Stanislav Nagy
Charles University, Faculty of Mathematics and Physics, Prague, Czech Rep.
e-mail: nagy@karlin.mff.cuni.cz

Robustness Aspects of Optimized Centroids

Jan Kalina and Patrik Janáček

Centroids are often used for object localization tasks, supervised segmentation in medical image analysis, or classification in other specific tasks. This paper starts by contributing to the theory of centroids by evaluating the effect of modified illumination on the weighted correlation coefficient. Further, robustness of various centroid-based tools is investigated in experiments related to mouth localization in non-standardized facial images or classification of high-dimensional data in a matched pairs design. The most robust results are obtained if the sparse centroid-based method for supervised learning is accompanied with an intrinsic variable selection. Robustness, sparsity, and energy-efficient computation turn out not to contradict the requirement on the optimal performance of the centroids.

Keywords: centroids, weighted correlation, robustness, contamination, centroid optimization

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic, e-mail: kalina@cs.cas.cz

Patrik Janáček
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic, e-mail: janacekpatrik@gmail.cz

Analysis of the Changes in the Polish Traditional Drugstores Market During COVID-19

Marcin Pelka, Antonio Irpino, and Michal Swachta

The COVID-19 pandemic has a significant impact on different aspects of economy. Usually the impact on traditonal economy. The paper presents the results of the analysis of the changes in the Polish traditional drugstores market during COVID-19 situation (from 2019 till 2021). Three different approaches with dynamic multidimensional scaling have been evaluated: classical data, symbolic interval-valued variables and symbolic histogram variables. The results show that symbolic historgram variables are able to capture all the variability in the data and refect all the changes.

Keywords: histogram variables, mutidimensional scaling, drugstores, covid-19

References

1. Bock H.-H., Diday E. (eds), Analysis of symbolic data. Exploratory statistics for complex. Springer, Heidelberg (2002)
2. de Carvalho, F. A. T.: Histograms in symbolic data analysis. Annals of Operations Research, 55(2), 299-322 (1995).
3. Irpino, A., & Verde, R.: A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In Data science and classification (pp. 185-192). Springer, Berlin, Heidelberg (2006).

Marcin Pelka
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: marcin.pelka@ue.wroc.pl

Antonio Irpino
Dipartimento di studi europei e mediterranei, Seconda Universitá degli Studi di Napoli, Caserta (CE), Italy, e-mail: antonio.irpino@unicampania.it

Michal Swachta
Wrocław University of Economics and Business, Komandorska 118/120, Wrocław, Poland
e-mail: 180459@student.ue.wroc.pl

Logistic Regression Models for Aggregated Data

Thomas Whitaker, Boris Beranger, and Scott Sisson

Logistic regression models are a popular and effective method to predict the probability of categorical response data. However, inference for these models can become computationally prohibitive for large datasets. Here we adapt ideas from symbolic data analysis to summarize the collection of predictor variables into histogram form, and perform inference on this summary dataset. We develop ideas based on composite likelihoods to derive an efficient one-versus-rest approximate composite likelihood model for histogram-based random variables, constructed from low-dimensional marginal histograms obtained from the full histogram. We demonstrate that this procedure can achieve comparable classification rates to the standard full data multinomial analysis and against state-of-the-art subsampling algorithms for logistic regression, but at a substantially lower computational cost. Performance is explored through simulated examples, and analyses of large supersymmetry and satellite crop classification datasets.

Keywords: class prediction, large datasets, one-versus-rest regression, random histograms, symbolic data analysis

Thomas Whitaker
UNSW Data Science Hub & School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia

Boris Beranger
UNSW Data Science Hub & School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia, e-mail: b.beranger@unsw.edu.au

Scott Sisson
UNSW Data Science Hub & School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia, e-mail: scott.sisson@unsw.edu.au

Nonparametric Regressions for Distributional Data

Albert Meco, Javier Arroyo, and Antonio Irpino

In distributional data individuals are described by means of distributions. While distributions are complex data representations, several methods have been successfully proposed to analyze data represented by distributional variables, giving rise to the field distributional data analysis [1]. In particular, several authors have proposed regression methods to model some specific kinds of linear relationships that can take place among distributional variables. However, such methods may suffer in cases where the linear relationship in the data is not the one the method can deal with, or when the relationship in the data is not linear. In such cases, nonparametric regression methods would be able to effectively model the underlying relationship.

We propose three nonparametric regression methods for distributional data: the kernel regression and the locally weighted regression using the Dias-Brito [2] and the Irpino-Verde [3] linear regressions for distributional data. The performance of the proposed methods and the linear approaches will be compared by means of a Monte Carlo experiment that will consider different underlying relationships in the data. In addition, we propose the use of several statistical measures and plots to analyze the regression fit and its errors to better understand how the regressions work.

Keywords: distributional data, nonparametric regression, locally-weighted learning

References

1. Brito, P., Dias, S. (eds.): Analysis of distributional data. Chapman and Hall/CRC, Boca Raton (2022)
2. Dias, S., Brito, P.: Linear regression model with histogram-valued variables. Stat. Anal. Data Min.. **8**: 75-113 (2015)
3. Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. Adv. Data Anal. Classif. **9**, 81–106 (2015).

Albert Meco
Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, Spain
e-mail: ameco@ucm.es

Javier Arroyo
Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Spain
e-mail: javier.arroyo@fdi.ucm.es

Antonio Irpino
Dipartimento di Matematica e Fisica, Università degli Studi della Campania "Luigi Vanvitelli", Italy
e-mail: antonio.irpino@unicampania.it

Hotspot Cluster Detection Based on Spatial Hierarchical Structure and its Software

Fumio Ishioka, Shoji Kajinishi, and Koji Kurihara

With the remarkable development of GIS in recent years, it has become easier to analyze and visualize a wide variety of geospatial data. Echelon analysis [3] was proposed as a method for objectively visualizing geospatial data by a topological hierarchical structure. An example application of using echelon analysis is hotspot detection, that is, detecting a subset of regions with significantly higher or lower relative risk in geospatial data. An *echelon scan method* we have proposed uses echelon's hierarchical structures to perform hotspot detection [2], and thereby has become possible to detect an arbitrary shaped cluster even when large amounts of data are targeted, which is difficult to detect by the conventional method. In this study, we introduce the R package for the echelon analysis and echelon scan method we have developed [1], and the web application that can execute these methods while performing them interactively. In addition, an example of analysis using actual data will be demonstrated.

Keywords: echelon analysis, echelon scan method, hotspot cluster

References

1. Ishioka, F.: echelon: The Echelon Analysis and the Detection of Spatial Clusters using Echelon Scan Method. R package version 0.1.0. (2020) https://cran.r-project.org/package=echelon
2. Kurihara, K., Ishioka, F., Kajinishi, S.: Spatial and temporal clustering based on the echelon scan technique and software analysis. Jpn. J. Stat. Data Sci. **3**, 313–332 (2020)
3. Myers, W.L., Patil, G.P., Joly, K.: Echelon approach to areas of concern in synoptic regional monitoring. Environmental and Ecological Statistics. **4**, 131–152 (1997)

Fumio Ishioka
Faculty of Environmental and Life Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, e-mail: fishioka@okayama-u.ac.jp

Shoji Kajinishi
Department of International Liberal Arts, Chugoku Gakuen University, 83 Niwase Okayama 701-0197, Japan, e-mail: skajinishi@g.cjc.ac.jp

Koji Kurihara
Faculty of Environmental and Life Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, e-mail: kurihara@okayama-u.ac.jp

Group Lasso Penalty for Spatially Clustered Coefficient Regression

Toshiki Sakai, Jun Tsuchida, and Hiroshi Yadohisa

Spatial data has variables with location information. One of the main purposes of spatial data analysis is to explain the relationships between objective variables and covariates according to this location information. Geographically weighted regression (GWR) [1] is a method that allows regression coefficients to vary by location. However, Li and Sang [3] showed that GWR can lead to numerical instability in estimations at locations with few surrounding observation points. They then proposed spatially clustered coefficients (SCC) regression using the fused lasso penalty. SCC enables the regression coefficients to be estimated with greater numerical stability than in GWR.
However, these methods do not assume groups among the covariates. The results hence tend to be difficult to interpret.
Herein, we propose a method that combines SCC and the group lasso penalty [2]. The proposed method makes it easier for regression coefficients to be the same at nearby locations while facilitating interpretation by selecting covariates on a group-by-group basis.

Keywords: geographically weighed regression, fused lasso, spatial data

References

1. Brunsdon et al.: Geographically weighted regression : a method for exploring spatial nonstationarity. Geographical Analysis **28**, 281–298 (1996)
2. Yuan, M. and Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49–67 (2006)
3. Li, F. and Sang, H.: Spatial homogeneity pursuit of regression coefficients for large datasets. Journal of the American Statistical Association. **114**, 1050–1062 (2019)

Toshiki Sakai
Graduate School of Culture and Information Science, Doshisha University, Tataramiyakodani 1-3, Kyotanabe City, Kyoto, Japan, e-mail: sakatarian711@gmail.com

Jun Tsuchida, Hiroshi Yadohisa
Department of Culture and Information Science, Doshisha University, Tataramiyakodani 1-3, Kyotanabe City, Kyoto, Japan

Visualization of the Number of New Positives for COVID-19 in Japan

Yoshiro Yamamoto, Sanetoshi Yamada, Mayumi Tanahashi, and Tadashi Imanishi

COVID-19 has been spreading in Japan and other parts of the world since 2020. In Japan, the number of newly confirmed positive cases is reported daily in the news. As the number of infected people increases, restraints on behavior and other measures have been taken to control the spread of infection. Our group has developed a system to visualize the number of positive cases of COVID-19 in five prefectures where university campuses are located, so that the status of the spread of COVID-19 infection can be monitored. ([1],[2], http://covid-map.bmi-tokai.jp/) The system provides a visualization of the number of newly confirmed positive cases in each municipality, and is designed to provide an interactive function to provide the user with the information he or she wants to know. We present the details of the data collection method as well as the visualization information realized by this system.

Acknowledgments This research was supported by a research grant (2021) from the Tokai University Union Supporting Association. We would like to express our deepest gratitude.

Keywords: COVID-19, visualization, interactive plot

References

1. Tanahashi,M., Yamada,S., Imanishi,T., Yamamoto,Y.: Choropleth map of newly infected people with COVID-19. 2021 19th International Conference on ICT and Knowledge Engineering. IEEE Xplore (2021)
2. Tanahashi,M., Yamamoto,Y.: Visualization of the distribution of newly infected persons with COVID 19 in the prefecture. 2020 18th International Conference on ICT and Knowledge Engineering. IEEE Xplore (2020)

Yoshiro Yamamoto
School of Science Tokai University, Kanagawa Japan, e-mail: yama@tokai-u.jp

Sanetoshi Yamada
School of Medicine Tokai University, Kanagawa Japan, e-mail: S.Yamada@star.tokai-u.jp

Mayumi Tanahashi
Graduate School of Science Tokai University, Kanagawa Japan,
e-mail: 0csfm004@hope.tokai-u.jp

Tadashi Imanishi
School of Medicine Tokai University, Kanagawa Japan, e-mail: imanishi@tokai.ac.jp

Two Simple but Efficient Elgorithms to Recognize Robinson Dissimilarities

Mikhaël Carmona, Guyslain Naves, Victor Chepoi, and Pascal Préa

A dissimilarity d on a set S of size n is said to be *Robinson* [3] if its matrix can be symmetrically permuted so that its elements do not decrease when moving away from the main diagonal along any row or column. Equivalently, S admits a total order $<$ such that $i < j < k$ implies that $d(i,j) \leq d(i,k)$ and $d(j,k) \leq d(i,k)$. Intuitively, d is Robinson if S can be represented by points on a line. Recognizing Robinson dissimilarities has numerous applications in seriation and classification. In this paper, we present two simple algorithms (inspired by Quicksort) to recognize Robinson dissimilarities. One of these algorithms uses partition refinement [2] and runs in $O(n^2 \log n)$, the other one uses PQ-trees [1] and runs in $O(n^3)$ in worst case and in $O(n^2)$ on average.

Keywords: Robinson dissimilarities, classification, seriation, pq-trees, partition refinement

Acknowledgements This work was supported in part by ANR project DISTANCIA (ANR-17-CE40-0015).

References

1. Booth, K.S., Lueker, G.S.: Testing for the consecutive ones property, interval graphs and graph planarity using PQ-tree algorithm, Journal of Computer and System Sciences **13**, 335–379 (1976).
2. Paige, R., Tarjan, R.E.: Three partition refinement algorithms, SIAM Journal on Computing **16**, 973–989 (1987).
3. Robinson, W.S.: A method for chronologically ordering archeological deposits. American Antiquity **16** , 293–301 (1951)

Mikhaël Carmona
LIS, Aix-Marseille Université, CNRS and Université de Toulon, Marseille, France and École Centrale Marseille, Marseille, France. e-mail: mikhael.carmona@lis-lab.fr

Victor Chepoi
LIS, Aix-Marseille Université, CNRS and Université de Toulon, Marseille, France.
e-mail: victor.chepoi@lis-lab.fr

Guyslain Naves
LIS, Aix-Marseille Université, CNRS and Université de Toulon, Marseille, France.
e-mail: guyslain.naves@lis-lab.fr

Pascal Préa
LIS, Aix-Marseille Université, CNRS and Université de Toulon, Marseille, France and École Centrale Marseille, Marseille, France. e-mail: pascal.prea@lis-lab.fr

Clustering with Missing Data: Which Imputation Model for Which Cluster Analysis Method?

Vincent Audigier, Ndèye Niang, and Matthieu Resche-Rigon

Multiple imputation (MI) is a popular method for dealing with missing values. One main advantage of MI is to dissociate the imputation phase and the analysis one. However, both are related since they are based on distribution assumptions that have to be consistent. This point is well known as *congeniality*.

In this talk, we discuss congeniality of imputation models and clustering on continuous data. First, we theoretically highlight how two joint modelling (JM) MI methods, using either general location model (JM-GL) or Dirichlet process mixture (JM-DP), could be congenial with various clustering methods. Then, we propose a new fully conditional specification (FCS) MI method with the same theoretical properties as JM-GL. Finally, we extend this FCS MI method from normal distribution to account for more complex distributions. Based on an extensive simulation study, all MI methods are compared for various cluster analysis methods (k-means, k-medoids, mixture model, hierarchical clustering).

This study highlights the partition accuracy is always improved when the imputation model accounts for clustered individuals. From this point of view, standard MI methods ignoring such a structure should be avoided. JM-GL and JM-DP should be recommended when data are distributed according to a Gaussian mixture model, while FCS methods outperform JM ones on data involving more complex distributions.

Keywords: clustering, missing data, multiple imputation, congeniality

Vincent Audigier
CNAM, CEDRIC-MSDMA, 2 rue Conté, 75003 Paris, France,
e-mail: vincent.audigier@cnam.fr

Ndèye Niang
CNAM, CEDRIC-MSDMA, 2 rue Conté, 75003 Paris, France,
e-mail: n-deye.niang_keita@cnam.fr

Matthieu Resche-Rigon
Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, 12 rue de la Grange-aux-Belles, Paris, 75010, France / Université de Paris, Paris, France / ECSTRRA, INSERM, UMR 1153, Paris, France,
e-mail: matthieu.resche-rigon@u-paris.fr

Hierarchies and Weak-hierarchies as Interval Convexities

Patrice Bertrand and Jean Diatta

There are several ways to characterize a hierarchy, one being a collection of nonempty subsets that are convex according to a type of interval function. This characterization in terms of interval convexity, extends to general classes of multilevel clusterings, thus providing a unifying theoretical framework [1, 2]. We expand this line of research, with a special attention to specifications allowing the capture of clusterings usually constructed in data mining practice, such as the Apresjan and the single-link hierarchies. We propose: (a) New characterizations of hierarchies and weak hierarchies as interval convexities, (b) Interval functions which induce known clustering schemes such as the Single Link hierarchy or the Apresjan hierarchy, (c) A sequence of nested families of interval convexities that is gradually increasing from the Apresjan hierarchy to the Single-Link hierarchy, which enables the detection of redundant clusters.

Keywords: weak hierarchy, interval convexity, single link hierarchy

References

1. Bertrand, P., Diatta, J.: Multilevel clustering models and interval convexities. Discrete Appl. Math. **222**, 54–66 (2017)
2. Changat, M., Narasimha-Shenoi, P.-G., Stadler, P.-F.: Axiomatic characterization of transit functions of weak hierarchies. Art Discrete Appl. Math. **2** #P1.01 (2019)
3. Yu, Z., Liu, W., Liu, W., Yang, Y., Li, M., Kumar, B. V. K. V.: On order-constrained transitive distance clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence **30(1)** (2016)

Patrice Bertrand
Ceremade, Université Paris-Dauphine-PSL, Pl. du Maréchal de Lattre de Tassigny, Paris, 75016, France, e-mail: patrice.bertrand@ceremade.dauphine.fr

Jean Diatta
LIM-EA2525, Université de La Réunion, Parc Technologique Universitaire, 2 rue Joseph Wetzell, Sainte Clotilde, 97490, France, e-mail: jean.diatta@univ-reunion.fr

A Rule-based Approach to Scoring Systems

Michael Rapp, Johannes Fürnkranz, and Eyke Hüllermeier

Scoring systems have a long history of active use in safety-critical domains such as healthcare and justice, where they provide guidance for making objective and accurate decisions. While scoring systems have often been handcrafted by domain experts in the past, the use of machine learning algorithms to deduce decision models from historic data is becoming more prevalent, also due to an increased availability of data and computational resources. In this work, we present an overview of existing methods for the data-driven construction of scoring systems and propose a taxonomy for characterizing the different types of models they produce. We further provide a new perspective on the topic by establishing a connection between scoring systems and additive rule models, and propose a rule-based methodology that allows for constructing scoring systems with different characteristics. In an experimental study, we investigate the effects of various constraints that are typically imposed on the complexity of scoring systems to facilitate their use by human practitioners. We also investigate how existing rule learning techniques help reduce negative impacts in terms of predictive accuracy they may entail.

Keywords: interpretable machine learning, scoring systems, rule learning

Michael Rapp
Chair of Machine Learning and Artificial Intelligence, Ludwig-Maximilians-Universität München, Munich, Germany, e-mail: michael.rapp@ifi.lmu.de

Johannes Fürnkranz
Institute for Application-oriented Knowledge Processing, Johannes Kepler University Linz, Linz, Austria, e-mail: juffi@faw.jku.at

Eyke Hüllermeier
Chair of Machine Learning and Artificial Intelligence, Ludwig-Maximilians-Universität München, Munich, Germany, e-mail: eyke@ifi.lmu.de

On Explaining Model Change Based on Feature Importance

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier

Understanding the decisions of complex machine learning (ML) models is vital for leveraging ML ethically and responsibly in everyday applications [1]. The research field of Explainable Artificial Intelligence (XAI) aims at increasing the interpretability of otherwise opaque ML systems. While XAI mainly focuses on static learning tasks, we are interested in explaining models in dynamic learning environments, such as online learning from real-time data streams, where models are trained in an incremental rather than a batch mode. Models in such dynamic settings need to react and adapt to changes in their environment. We motivate the problem of explaining these dynamic models by directly explaining the model change, i.e., the difference between models before and after adaptation, instead of the models per se. We discuss how this problem may be approached by agnostic explanation methods such as Feature Importance (FI) and, more specifically, an adaption of the well-known Permutation Feature Importance (PFI) [2]. We present an incremental version of PFI and showcase how existing algorithms for detecting changes in data streams can be adapted to explain model change directly.

Keywords: explainable artificial intelligence, explaining model change, incremental learning, permutation feature importance

Acknowledgements We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021–438445824.

References

1. Adadi, A. and Berrada M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. **6**, 52138–52160 (2018)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32. (2001)

Maximilian Muschalik
LMU Munich, Germany, e-mail: maximilian.muschalik@ifi.lmu.de

Fabian Fumagalli
Bielefeld University, Germany, e-mail: ffumagalli@techfak.uni-bielefeld.de

Prof. Dr. Barbara Hammer
Bielefeld University, Germany, e-mail: bhammer@techfak.uni-bielefeld.de

Prof. Dr. Eyke Hüllermeier
LMU Munich, Germany, e-mail: eyke@ifi.lmu.de

Interpretable Multi-class Trees for Travel Choice Mode Analysis

Christian Riccio, Andrea Papola, Michele Staiano, and Roberta Siciliano

Analysis of travel mode choice is fundamental to forecast travel demand when planning intervention on the supply system. Commonly, this is conducted via Random Utilities Models (RUMs) which relies on the random utility theory [1]. Recently, the large availability of travel demand data, mainly from smartphones, has increasingly led to the use of machine learning models that find their ideal context of use in big data. Although such models are generally capable of good performance, their intrinsic black-box nature is a critical aspect. Hence, to fruitfully apply Machine Learning approaches to travel mode choice, some enhancements must be considered. Recently, interpretable machine learning approach [2] has been proposed. The main idea is to improve the output results of random forests with some additional aids. This paper provides a new framework to approach interpretable machine learning using tree-based methods in combination with classical models. The basic rationale with the main theoretical aspects can be found in old papers [3, 4, 5]. A fresh approach integrating logistic regression and latent budget models will be demonstrated for a multi-class response problem typical in transportation studies.

Keywords: tree-based methods, logistic regression, travel mode choice

References

1. Cascetta, E., Papola, A.: Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. Transport. Res. C-Emer. **9**(4), 249–263 (2001)
2. Kim, E-J.: Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach. J. Adv. Transp. **2021**, 6685004 (2021)
3. Siciliano, R., Mola, F.: Multivariate data analysis and modeling through classification and regression trees. Comput. Stat. & Data An. **2021**, 6685004 (2021)
4. Siciliano, R.: Latent budget trees for multiple classification. I In: Vichi, M., Opitz, O. (eds) Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg (1999)
5. Mola, F., Klaschka, J., Siciliano, R.: Logistic classification trees. In: Prat, A. (ed.) COMPSTAT 1996, pp. 373-378. Physica-Verlag, Heidelberg (1996)

Christian Riccio · Andrea Papola
DICEA, 21, Via Claudio - 80125 Napoli, Italy,
e-mail: {christian.riccio,andrea.papola}@unina.it

Michele Staiano
DII, 80, P.le Tecchio - 80125 Napoli, Italy, e-mail: michele.staiano@unina.it

Roberta Siciliano
DIETI, 21, Via Claudio - 80125 Napoli, Italy, e-mail: roberta@unina.it

Identification of Driver Genes in Glioblastoma via Regularized Classification

Marta Belchior Lopes and Susana Vinga

Tumor heterogeneity is a major driver of tumor progression and treatment failure. In the particular case of glioblastoma (GBM), the most common and aggressive primary brain malignancy, intratumoral molecular heterogeneity translates into different tumor cell clones in a single patient, with different selective advantages, which makes available therapy options ineffective. The identification of tumor molecular changes at the cell level is a key to understanding tumor heterogeneity and providing insights into the development of novel targeted therapies. With the rise of omics technologies, it is now possible to extract from a single cell a huge amount of information related to the cell functioning (e.g., genomics, transcriptomics, proteomics). However, these data came at the cost of high dimensionality (the number of features greatly outnumbering the number of observations), which requires appropriate statistical and machine learning tools to extract relevant information from these complex molecular networks. In this work, a strategy based on regularized logistic regression with network information is applied to single-cell RNA sequencing (RNA-seq) data from GBM patients for classifying cells into distinct neoplastic cell groups and normal cells, while selecting the features discriminating between the classes as putative GBM therapy targets. The relevance of the extracted features is supported by literature reports on their established role in GBM, their significance in the survival outcomes in bulk GBM RNA-Seq data, and their association with several Gene Ontology biological process terms.

Keywords: glioblastoma, transcriptomics, regularized classification, high dimentionality, network

Marta Belchior Lopes
Center for Mathematics and Applications (CMA) and NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), NOVA School of Science and Technology (FCT NOVA),
e-mail: marta.lopes@fct.unl.pt

Susana Vinga
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa,
e-mail: susanavinga@tecnico.ulisboa.pt

Outlier Detection: a Procedure to Capture Atypical Groups of Observations

Ana Helena Tavares, Vera Afreixo, and Paula Brito

In this work, we introduce the concept of *atypical group* of observations and propose a procedure for its identification. By atypical group, we mean a cluster of observations whose 'mean' pattern stands out from the majority of the 'mean' patterns of the remaining clusters. Challenges that arise in atypical group detection are firstly to identify a meaningful segmentation of the data, and secondly to flag the atypical segments. Our work focus on data whose elements are discrete distributions.

If heterogeneous datasets, where distinct patterns coexist, can validly be clustered, then the class prototypes provide a simplified description of data. Thus, the key idea of our proposal is to combine a clustering method with a functional outlyingness criterion to capture atypical class prototypes.

To identify a segmentation of the distributional data we iteratively combine two steps. The first creates a hierarchy of clusters, while the second flags atypical curves within each cluster, based on a measure of functional outlyingness which accounts for the shape of the distributions [1]. Segments with atypical curves, are forwarded for (sub)clustering, and the procedure is repeated until no outlying curves are identified in clusters. Once the final partition is obtained, each cluster is represented by a class prototype, whose outlyingness is evaluated according to the same functional approach. Clusters with an atypical class prototype are pointed as atypical.

We apply our procedure to investigate clusters of genomic words in human DNA by studying their inter-word lag distributions. These experiments demonstrate the potential of the new method for identifying clusters of words with outlying patterns.

Keywords: outlyingness, clustering, distributional data, functional data

References

1. Rousseeuw, P. J., Raymaekers, J., Hubert, M.: A measure of directional outlyingness with applications to image data and video. J Comput Graph Stat. **27:2**, 345-359 (2018)

Ana Helena Tavares
Águeda School of Technology and Managment, University of Aveiro & CIDMA, Portugal,
e-mail: ahtavares@ua.pt

Vera Afreixo
Department of Mathematics, University of Aveiro & CIDMA, Portugal, e-mail: vera@ua.pt

Paula Brito
FEP, University of Porto & LIAAD INESC TEC, Portugal, e-mail: mpbrito@fep.up.pt

Bayesian Classification and Non-Bayesian Label Estimation via EM Algorithm to Identify Differential Expression in Omics Data: a Comparative Study

Marilia Antunes and Lisete Sousa

Analyses of omics data (biomolecules including large macromolecules such as proteins and nucleic acids, as well as small molecules such as metabolites and natural products) allow comprehensive genome-wide analysis of complex diseases, offering a major advantage over previous candidate gene analysis or pathway analysis. One of the main goals in high-throughput data analysis is the identification of biomolecules among several thousands that differ between different sample groups.
Omics data classification problem is studied considering the ratio of the expression levels and a non-observed categorical variable indicating how differentially expressed each biomolecule is: *non differentially expressed*, *down-regulated* or *up-regulated*. Assuming that the ratios follow a mixture of gamma distributions, two methods are proposed [1]. The first method is based on a hierarchical Bayesian model. The conditional probability of a biomolecule to belong to each group is calculated and the biomolecule is assigned to the group for which this conditional probability is higher. The second method uses the EM algorithm to estimate the most likely group label for each biomolecule, that is, to assign the biomolecule to the group which contains it with the higher estimated probability. Both approaches are applied to omics data and results are compared.

Keywords: differential expression, EM algorithm, hierarchical Bayesian model, mixture models

References

1. Antunes, M., Sousa, L.: Bayesian classification and non-Bayesian label estimation via EM algorithm to identify differentially expressed genes: a comparative study. Biometrical J. **50**(5), 824–836 (2008)

Marilia Antunes
CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal e-mail: marilia.antunes@ciencias.ulisboa.pt

Lisete Sousa
CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, e-mail: lmsousa@ciencias.ulisboa.pt

Sequence-aware Item Recommendations for Multiply Repeated User-item Interactions

Juan Pablo Equihua, Maged Ali, Henrik Nordmark, and Berthold Lausen

Recommender systems are one of the most successful applications of machine learning in a wide variety of application domains, such as e-commerce, media streaming, email marketing, and virtually every industry where personalisation facilitates better user experience or boosts customer engagement [1]. Their main goal is analysing past users' behaviour to predict which items may be of interest for users, and are typically built with the use of matrix-completion techniques such as collaborative filtering. However, although these approaches have achieved tremendous success in several applications, their effectiveness is still limited when users might interact with the same items multiple times, or when user preferences change over time [2].

Inspired by Natural Language Processing techniques to process, and analyse sequences of text, we designed a recommender system that accounts to the order of user-item interactions in a sequential framework to make recommendations [3]. This method is empirically shown to predict accurate probabilities of future user-item interactions in retail environments, and outperform matrix-completion and similar sequential approaches by increasing sales up to 130% in different A/B tests.

Keywords: recommender systems, natural language processing, deep learning.

References

1. Massimo Q., Paolo C., Dietmar J.: Sequence-Aware Recommender Systems. ACM Computing Surveys (2019)
2. Thang D., Neil V., Brian K.: Generating Realistic Sequences of Customer-Level Transactions for Retail Datasets. IEEE International Conference on Data Mining Workshops. 820-827 (2018)
3. Young-Jun K., Lucas M., Matthias G.: Collaborative Recurrent Neural Networks for Dynamic Recommender Systems. ACML (2016)

Juan Pablo Equihua
Mathematical Sciences, University of Essex, Colchester, UK and Profusion Ltd., London, UK, e-mail: je18890@essex.ac.uk

Maged Ali
Essex Business School, University of Essex, Colchester, UK, e-mail: maaali@essex.ac.uk

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

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

High-dimensional Linear Regression Estimation

Mauro Iannuzzi and Matteo Farnè

The least square solution for the estimation of the parameters in a multiple linear regression model is not unique when the number of variables is larger than the number of units and can be very inaccurate when the number of units is larger than the number of variables, but the covariate space is high-dimensional.

In [1] a new estimator for large covariance matrices is proposed. The method is called UNALCE (UNshrunk ALgebric Covariance Estimator) and is based on the decomposition of the high-dimensional covariance matrix into the sum of a low rank (L) and a sparse (S) component.

In this poster the effect of different covariance estimators on the statistical properties of the estimates of the regression coefficients in a multiple linear regression model is assessed through a wide simulation study.

The goal is to test whether, if appropriately optimized, UNALCE is able to increase the accuracy in estimation, given a high-dimensional context. The results are compared with the standard least squares one (when feasible), with the RIDGE [2] and with ALCE (a variant of UNALCE that stops before the unshrinkage step).

Keywords: multiple linear regression, high dimensions, nuclear norm

References

1. Farné, M., Montanari, A.: A large covariance matrix estimator under intermediate spikiness regimes. Journal of Multivariate Analysis **176**, 104577 (2020)
2. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

Mauro Iannuzzi
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41,
e-mail: mauro.iannnuzzi@studio.unibo.it

Matteo Farnè
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41,
e-mail: matteo.farne@unibo.it

Experimental Study of Similarity Measures for Clustering Uncertain Time Series

Michael Dinzinger, Michael Franklin Mbouopda, and Engelbert Mephu Nguifo

Uncertain time series (uTS) are time series whose values are not precisely known. Each value in such time seris can be given as a best estimate and an error deviation on that estimate. These kind of time series are preponderant in transient astrophysics where transient objects are characterized by the time series of their light curves which are uncertain because of many factors including moonlight, twilight and atmospheric factors. An example of uTS dataset can be found at `https://www.kaggle.com/c/PLAsTiCC-2018`. Similarly to traditional time series, machine learning can be used to analyze uTS. This analyzis is generally performed in the literature using uncertain similarity measures. In particular, uTS clustering has been performed using FOTS, an uncertain similarity measure based on eigenvalues decomposition [1]. Elsewhere, the uncertain euclidean distance (UED), which is based on uncertainty propagation has been proposed and used to perform the classification of uTS [2]. Given UED performance on supervised classication, the goal of this work is to assess the effectiveness of this uncertain measure for uTS clustering.

A preliminary experiment has been conducted in that direction, the source code and results of the experiment are publicly available online[1]. In the experiment, FOTS, UED and euclidean distance are compared as measures for uTS clustering using the datasets from [2]. The obtained results revealed that UED is a promising uncertain measure for uTS clustering. As future direction, an extended experiment with other uncertain similarity measures such as DUST and PROUD [3] will be conducted.

Keywords: time series, clustering, uncertainty, similarity

Acknowledgements This work has been partially supported by the LabEx IMobS3.

References

1. Siyou Fotso, V. S., Mephu Nguifo, E., and Vaslin, P.: Frobenius correlation based u-shapelets discovery for time series clustering. Pattern Recognition, 2020, vol. **103**, p. 107301.
2. Mbouopda, M. F. and Mephu Nguifo, E.: Uncertain time series classification with shapelet transform. In: International Conference on Data Mining Workshops. IEEE, 2020. p. 259-266.
3. Dallachiesa, M., Nushi, B., Mirylenka, K., and Palpanas, T.: Uncertain timeseries Similarity: Return to the Basics. In: VLDB Endowment, 2012, vol. **11**, p. 1662–1673.

Michael Dinzinger, Michael Franklin Mbouopda and Engelbert Mephu Nguifo
University Clermont Auvergne, Clermont Auvergne INP, LIMOS, ISIMA, France
e-mail: `michael.dinzinger@etu.uca.fr`,
e-mail: `{michael.mbouopda,engelbert.mephu_nguifo}@uca.fr`

[1] `https://github.com/dim35216/UED_Clustering.git`

Assessing the Status of Two Data-limited Skates Landed in Portuguese Ports Using an Empirical Catch Rule

Erick Chatalov, Ivone Figueiredo, Lisete Sousa, and Bárbara Pereira

Worldwide there are concerns on the impact of fisheries on marine ecosystems. The goal 14 of the UN sustainable development aims to Conserve and Sustainably use the Oceans, Seas and Marine Resources. To reach this goal it is required that fisheries be adequately managed. Despite that, for many fisheries the data available is deficient and do not allow to have reliable estimates of stock status. In those cases, international agreements request the adoption and implementation of the Precautionary Approach. *Raja brachyura* and *Raja clavata* are two elasmobranch species characterized by life-cycle traits that make them particularly vulnerable to fishing. In Portuguese waters data available for these species are limited despite their relative importance in Rajidae landings, being mainly caught by fishing vessels operating bottom trawl and trammel nets. This study applies the methodology proposed by Fischer et al. [1] for evaluating the performance of a modified "2 over 3" rule as the management procedure, after a 100-year fishing history simulated from age-structured operating models for data-limited stocks, using the available knowledge on biological parameters of the two species. A genetic algorithm was applied to improve the performance of a data-limited catch rule. Different levels of uncertainty on the input data were explored and its impact on the estimation evaluated.

Keywords: empirical catch rule, genetic algorithm, age-structured operating models, Rajidae species

References

1. Fischer, S.H., De Oliveira, J.A., Mumford, J.D., Kell, L.T.: Using a genetic algorithm to optimize a data-limited catch rule. ICES J. Mar. Sci., **78**(4), 1311–1323 (2021)

Erick Chatalov
FCUL, Universidade de Lisboa, e-mail: erickchatalov@gmail.com

Ivone Figueiredo
Instituto Português do Mar e da Atmosfera, and CEAUL, e-mail: ifigueiredo@ipma.pt

Lisete Sousa
FCUL and CEAUL, Universidade de Lisboa, e-mail: lmsousa@ciencias.ulisboa.pt

Bárbara Pereira
Instituto Português do Mar e da Atmosfera, e-mail: bpereira@ipma.pt

Machine Learning Approach to Identify Factors that Influence Accident Severity

Daniel Santos, Vitor Nogueira, José Saias, Paulo Quaresma, Paulo Infante, Gonçalo Jacinto, Anabela Afonso, Leonor Rego, Pedro Nogueira, Marcelo Silva, Rosalina Pisco Costa, Patrícia Gois, and Paulo Rebelo Manuel

Since the twentieth century, road traffic accidents became a severe public health concern, with deaths and injuries posing a serious threat to world health and a negative influence on social and economic progress. One of the primary goals of accident data analysis is to determine the main factors that contribute to a traffic accident. This study aims to create a Machine Learning approach capable of identifying the factors that influence accident severity (seriously injured/dead or lightly injured/no injured), supporting the analysis of accident data. A four-year traffic accident data set from 2016 to 2019 in the Portuguese district of Setúbal is used. Clustering, Random Forests and C5.0 rule models are some of the techniques used to select the most influential factors and represent them in rule sets. Results show that a rule-based model using the C5.0 algorithm can accurately identify the most relevant factors describing road accident severity. Factors such as accidents involving motorcycles and pedestrian running over are the most prominent factors in our data.

Keywords: machine learning, road accident data, rule-based model

Daniel Santos and Leonor Rego
Universidade de Évora, e-mail: dfsantos@uevora.pt,lrego@uevora.pt

Vitor Nogueira and José Saias and Paulo Quaresma
Algoritmi Research Centre, Depto. de Informática, Universidade de Évora,
e-mail: vbn@uevora.pt,jsaias@uevora.pt,pq@uevora.pt

Paulo Infante, Gonçalo Jacinto and Anabela Afonso
CIMA, Depto. de Matemática,Universidade de Évora,
e-mail: pinfante@uevora.pt,gjcj@uevora.pt,aafonso@uevora.pt

Pedro Nogueira and Marcelo Silva
ICT, Depto. de Geociências,Universidade de Évora,
e-mail: pmn@uevora.pt,marcelogs@uevora.pt

Rosalina Costa
CICS.NOVA.UEVORA, Depto. de Sociologia, Universidade de Évora,
e-mail: rosalina@uevora.pt

Patrícia Gois
Depto. de Artes Visuais e Design/EA, University of Évora, e-mail: pafg@uevora.pt

Paulo Rebelo Manuel
CIMA, Universidade de Évora, e-mail: pjsrm@uevora.pt

Trade and Bank Credit of Portuguese SMEs: a Panel Data Application

Carla Henriques, Pedro Pinto, and Carolina Cardoso

Financial constraints are an obstacle to the growth and development of small and medium-sized enterprises (SME) [3]. The literature review shows that SMEs preferentially resort to bank credit to face these constraints [2]. However, given the limitations to which they are subject and the constraints they encounter with the banking system, trade credit appears as an alternative for many companies [2]. To study the determinants of bank credit and trade credit, regression models with panel data were used, considering a time horizon of ten years (2010 to 2019). The data were collected through the SABI (Iberian Balance Sheets Analysis System), selecting 5860 Portuguese SMEs with a total of 58,600 panel data records. The models include the firm's information on the return on assets, collateral security, current ratio, turnover growth and Altman's Z score of bankruptcy prediction [1], as independent variables. The time dummies were also included. Several models were estimated using fixed effects and random effects estimation and the same conclusions were drawn. This study is considered relevant for Portuguese SMEs, banking institutions and stakeholders, as the main factors that influence the use of bank credit and trade credit are analyzed along with an indicator of financial distress.

Keywords: panel data, regression models, trade credit, bank credit

References

1. Altman, E. I.:Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy (Frontiers in Finance Series). John Wiley & Sons (1983)
2. D'Ignazio, A., Menon, C.: Causal Effect of Credit Guarantees for Small- and Medium-Sized Enterprises: Evidence from Italy. The Scandinavian Journal of Ecnonomics, **122(1)**, 191-218. (2020)
3. Moscalu, M., Girardone, C., Calabrese, R.: SMEs' growth under financing constrains and banking markets integration in the euro area. Journal of Small Business Management. **58(4)**, 707–746 (2020)

Carla Henriques
CMUC, Coimbra, and Polytechnic Institute of Viseu, Viseu, e-mail: carlahenriq@estgv.ipv.pt

Pedro Pinto
CISeD, Viseu, and Polytechnic Institute of Viseu, Viseu, e-mail: spinto@estgv.ipv.pt

Carolina Cardoso
Polytechnic Institute of Viseu, Viseu, e-mail: carolinaesteves96@outlook.com

Hausdorff Distance: a Powerful Tool for Matching Households and Individuals in Historical Censuses

Thais Pacheco Menezes, Michael Fop, and Thomas Brendan Murphy

Matching households and individuals in historical censuses can be difficult due to the absence of unique identifiers, typographical errors, and changes in attributes over time. The tools of record linkage are of great assistance when linking households and individuals in historical censuses. In this work, we define a general multi-step record linkage procedure that allows the incorporation of household information to improve the process of matching entities across different databases. We propose using the Hausdorff distance when comparing households in historical censuses. A constrained logistic regression model with attribute level Hausdorff distances is developed to estimate the probability of a match between any two households. The probabilities from this model are then employed to match households across the databases. Subsequently, individuals within households are matched using a logistic regression based on attribute level distances. The probabilities estimated from this regression are used in a linear programming optimization framework to enforce one-to-one matches between individuals in the matched household across the databases. The methodology is developed in application to record linkage of the Irish census databases of 1901 and 1911. The analysis focuses on a number of regions for which labels of matching households and individuals are available, allowing training and testing of the procedure. The approach is shown to yield 65.15% correct household matches for regions that are close to the region used for model training and 58.70% correct household matches for more distant regions. When matching individuals within households, an average correct individual match rate of 86.2% is found for individuals within correctly matching households.

Keywords: census, Hausdorff distance, matching databases, record linkage

Thais Pacheco Menezes
School of Mathematics and Statistics, University College Dublin (UCD), Belfield, Dublin 4 - Ireland, e-mail: thais.pachecomenezes@ucdconnect.ie

Michael Fop
School of Mathematics and Statistics, University College Dublin (UCD), Belfield, Dublin 4 - Ireland, e-mail: michael.fop@ucd.ie

Thomas Brendan Murphy
School of Mathematics and Statistics, University College Dublin (UCD), Belfield, Dublin 4 - Ireland, e-mail: brendan.murphy@ucd.ie

Model-based Tri-clustering

Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif

Classical clustering procedures seek to separately construct an optimal partition of rows or columns or sometimes of rows and columns simultaneously. In this latter, co-clustering methods organize the data into homogeneous blocks. Methods of this kind have practical importance in a wide variety of applications. However, tensor data representation is a handy tool to represent data with complex structures. The three-way tensors are used in different fields like recommender systems, medical fields and social studies. Thereby, extending co-clustering to tri-clustering is a good manner to harness this kind of data. Several tri-clustering algorithms have been proposed in the literature. As suggested in [1] and through our investigations, we propose to classify the existing tri-clustering methods into five families: stochastic, greedy, genetic, tensor factorization and co-clustering based approaches. In our proposal we propose a flexible model-based tri-clustering. In order to use the model in a clustering setting, we want to jointly infer the latent variables $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{S}$ and learning the model parameters. We develop an approach based on Variational Expectation-Maximization and we derive, thereby, effective tri-clustering algorithms capable to reveal homogeneous sub-tensors from a 3-way tensor sparse or not. We illustrate the performances of these algorithms through numerical experiments on simulated and real-case datasets comparing with baseline algorithms [2, 3] in different fields including text-mining.

Keywords: tri-clustering, tensor, data science, text-mining

References

1. Henriques, R., d Madeira, S. C.: Triclustering algorithms for three-dimensional data analysis: a comprehensive survey. In: ACM Comput. Surv. **51**(5), 1-43 (2018)
2. Guigourès, R., Boullé, M., Rossi, F.: Discovering patterns in time-varying graphs: a triclustering approach. In: Advances in Data Analysis and Classification. **12**(3), 509-536 (2018)
3. Boutalbi, R., Labiod, L., and Nadif, M. Tensorclus: A python library for tensor (co)-clustering. In: Neurocomputing, pp. 464–468, (2022)

Rafika Boutalbi
Institute for Parallel and Distributed Systems, Analytic Computing, University of Stuttgart, Germany, e-mail: rafika.boutalbi@ipvs.uni-stuttgart.de

Lazhar Labiod and Mohamed Nadif
Centre Borelli UMR 9010, Université Paris Cité, France,
e-mail: lazhar.labiod@u-paris.fr,mohamed.nadif@u-paris.fr

Analyzing the Effects of Deviations from Normality on the Latent Growth Curve Models Goodness-of-fit

Catarina Marques, Maria de Fátima Salgueiro, and Paula C.R. Vicente

Latent growth curve models (LGCM) became in recent years a very popular technique for longitudinal data analysis: they allow individuals to have distinct growth trajectories over time [1]. Although the LGCM specified model structure imposes normality assumptions, the data analyst often faces data deviations from normality, implying mild, moderate or even severe values for skewness and or kurtosis. In the current research, a Monte Carlo simulation study was conducted in order to investigate the effect of observed data deviations from normality on goodness-of-fit indices. A new approach to generate multivariate non-normal distributed data was used: the VITA method [2]. This method is a covariance model simulation method using regular vines. The dependency structure is determined by bivariate copulae and a nested set of trees.

One thousand datasets were randomly generated from regular vines using Clayton copula and three marginal distributions (Normal, Student 3 and Gamma). The multivariate normal distribution was also used for data generation. LGCM with unconditional linear growth was considered. Three time points and distinct combinations of sample sizes were used. The impacts of such deviations on goodness-of-fit measures are discussed.

Keywords: goodness-of-fit indices; LGCM, non-normality data, VITA method.

References

1. Bollen, K.A., Curran, P.J.: Latent Curve Models - A Structural Equation Perspective. John Wiley & Sons, New Jersey, USA. (2006)
2. Grønneberg, S., Foldnes, N.: Covariance Model Simulation using Regular Vines. Psychometrika **82(4)**, 1035 – 1051 (2017).

Catarina Marques
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Lisboa, Portugal,
e-mail: catarina.marques@iscte-iul.pt

Maria de Fátima Salgueiro
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Lisboa, Portugal,
e-mail: fatima.salgueiro@iscte-iul.pt

Paula C.R. Vicente
Universidade Lusófona de Humanidades e Tecnologias, Lisboa, Portugal,
e-mail: pvicente@ulusofona.pt

Transformation Mixture Modeling for Skewed Data Groups with Heavy Tails

Yana Melnykov, Xuwen Zhu, and Volodymyr Melnykov

Gaussian mixture models have been the most popular mixtures in literature over years. However, the validity of the normality assumption for individual groups is often violated. In such cases, distributions that can model skewness as well as heavy tail behavior are often chosen as mixture components. The proposed contaminated transformation mixture model employs the idea of transformations to symmetry and can effectively model skewness, heavy tails, and scatter in data. A model selection algorithm is proposed to find models that perform well in terms of the data fit but involve fewer parameters. The utility of the methodology is illustrated on simulated and classification data sets.

Keywords: finite mixture model, cluster analysis, transformation, skewness

References

1. Melnykov, Y., Zhu, X., and Melnykov, V.: Transformation Mixture Modeling for Skewed Data Groups with Heavy Tails and Scatter. Computational Statistics **36**, 61–78 (2021)
2. Zhu, X. and Melnykov, V.: Manly Transformation in Finite Mixture Modeling. Computational Statistics and Data Analysis **121**, 190-208 (2018)

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

Xuwen Zhu
The University of Alabama, USA, e-mail: xzhu20@cba.ua.edu

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

A Simulation Study on Variable Selection in Mixture Regression Models

Susana Faria

Finite mixture regression models provide a flexible tool for modeling data that arise from a heterogeneous population, where the relationship between the response and the covariates varies across the sub-populations [3]. In the applications of these models, a large number of covariates are often used and their contributions toward the response variable vary from one component to another of the mixture model. For this reason, variable selection assumes a great relevance for mixture models, something particularly notorious in the last few years.

Variable selection via penalized likelihood has attracted great attention in recent literature. In particular, [2] have investigated the variable selection problem for finite mixture regression models with versions of the penalty functions.

In this work we analyze the problem of variable selection in finite mixture regression models in the presence of a large number of explanatory variables. We study the performance of a penalized likelihood approach in identifying the most relevant subset of covariates using the Classification Expectation-Maximization algorithm (CEM) [1].

Simulation examples indicate that the method works well in terms of both variable selection and prediction accuracy.

Keywords: mixture of linear models, variable selection, penalized likelihood, classification expectation-maximization (CEM) algorithm

Acknowledgements Supported by Portuguese funds through the Portuguese Foundation for Science and Technology, within project PTDC/MAT-STA/28243/2017.

References

1. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis **14(3)**, 315-–332 (1992).
2. Khalili A., Chen J.: Variable selection in finite mixture of regression models. J. Amer. Statist. Assoc. **102**, 1025–1038 (2007).
3. McLachlan G.J., Peel D.: Finite Mixture Models. Wiley, New York (2000).

Susana Faria
University of Minho, Centre of Molecular and Environmental Biology and Department of Mathematics, Guimarães, Portugal, e-mail: sfaria@math.uminho.pt

Hybrid Forecasting Combinations by Feature Based Metalearning

Moises Santo, Andre C.P.L.F. de Carvalho, and Carlos Soares

A challenge in time series forecasting is selecting the technique that will induce the best forecasting model for a given time series. Metalearning is a good alternative to reduce the computational cost and attend to the high need for specialized knowledge in this area [1]. In recent works, model selection frameworks successfully applied metalearning to select time series forecasting methods. Allied to metalearning, combining forecasts is an up-and-coming alternative to investigate. One of the main results of a time series competition to assess time series forecasting models, the Makridakis M4 competition was the victory of a hybrid method that combined the forecasts of statistical and machine learning models [3]. Several time series forecasting approaches have applied Seasonal-Trend decomposition based on Loess (STL) and Empirical Mode Decomposition (EMD) to divide the time series into components with different properties to generate hybrid forecasting models. However, the models for each component have usually been selected arbitrarily. This work investigates the automatic selection of promising hybrid model combinations for univariate time series forecasting by using feature-based metalearning. According to previous experiments [2], under certain conditions, the application of the STL decomposition method for the formation of additive hybrid combinations presents better results than the use of individual models.

Keywords: time series, hybrid forecasting, metalearning

References

1. Brazdil, P., Carrier, C.G., Soares, C, Vilalta, R.: Metalearning: Applications to data mining. Springer Science & Business Media (2008)
2. Silvestre, G.D., Santos, M.R., Carvalho, A.C.P.L.F.: Seasonal-Trend decomposition based on Loess+ Machine Learning: Hybrid Forecasting for Monthly Univariate Time Series. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-7. IEEE (2021)
3. Smyl, S.: A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting **36**, 75-85 (2020)

Moises Santos
University of São Paulo, São Carlos, Brazil, e-mail: mmrsantos@usp.br

Andre C.P.L.F. de Carvalho
University of São Paulo, São Carlos, Brazil, e-mail: andre@icmc.usp.br

Carlos Soares
Faculty of Engineering - University of Porto, Porto, Portugal, e-mail: csoares@fe.up.pt

On the Measurement of Household Subjective Poverty: Concepts and Application

Aleksandra Łuczak and Sławomir Kalinowski

Poverty is a multi-dimensional phenomenon that cannot be directly measured correctly by a single indicator. Research on poverty uses both objective and subjective indicators. Objective measures do not show the complete nature of poverty as they only examine economic conditions, mainly household income or expenses, or their basic needs. Hence, our research focuses on subjective poverty, which includes the diversity of perceptions of poverty among respondents. The aim of the research is to present and compare methodological approaches to the construction of a synthetic measure of subjective household poverty. The research took into account the aggregation of variables describing the past, present, and future. The procedure of constructing a synthetic measure utilized the measure of distance between triangular fuzzy numbers [2] and the generalized distance measure [5]. However, the aggregation of variables relied on three fuzzy methods based on the ideas of Hellwig [1], TOPSIS [4] and Chen [1]. The approaches were used to assess the level of subjective household poverty in Poland. The study was based on data collected using the CAWI method in April 2021. The use of fuzzy approaches to the assessment of subjective poverty allows more precise determination of its level than with the standard measurement. In addition, the subjective poverty measure can also be used as a basis for formulating poverty reduction strategies.

Keywords: subjective poverty, fuzzy hellwig's method, fuzzy topsis, gdm

References

1. Chen, S.M.: Evaluating weapon systems using fuzzy arithmetic operations, Fuzzy Set Syst 77 (3), pp. 265–276 (1996)
2. Chen, C.T. Extensions of the TOPSIS for group decision-making under fuzzy environment. Fuzzy Set Syst, 114, 1–9 (2000)
3. Hellwig, Z.: Zastosowania metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę wykwalifikowanych kadr. Statistical Review 4, 307–327 (1968)
4. Hwang, C.L.; Yoon, K.: Multiple attribute decision making: methods and applications. Springer-Verlag, New York (1981)
5. Walesiak, M.: Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R. Wyd. Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław (2011)

Aleksandra Łuczak
Poznań University of Life Sciences, Faculty of Economics, Poznań, Poland,
e-mail: aleksandra.luczak@up.poznan.pl

Sławomir Kalinowski
Polish Academy of Sciences, Institute of Rural and Agricultural Development, Warsaw, Poland,
e-mail: skalinowski@irwirpan.waw.pl

Socio-economic Classification of Territorial Units: Extreme Value Theory-based Methods as Support for the Construction of a Synthetic Index

Aleksandra Łuczak and Małgorzata Just

A socio-economic classification of territorial units is helpful to assess their condition and plan their development. However, this process reveals many problems, such as data availability, selection of indicators and their measurement, selection of appropriate methods for normalization, weighing and aggregation of indicators. Aggregation of data depend on kind of data and distribution of variables. One of the main problems for real data are atypical observations. They are significant complication in the analysis of complex economic phenomena, because they have a strong influence on the obtained results [3]. The main goal is to present a comprehensive linear ordering procedure using the positional version of TOPSIS and methods of extreme value theory to assess the level of socio-economic development of various types of territorial units. We propose a two-step procedure based on six automatic methods that identify the tail of the variable distribution (atypical observations). Moreover, in order to eliminate the impact of asymmetry, mainly in the central part of the variable distribution, we also use the positional TOPSIS method based on ideas of Hellwig [1], as well as Hwang and Yoon [2]. The procedure proposed was used to assess the socio-economic situation of voivodships, districts and municipalities in Poland in 2019. The conducted research allows to expand the spectrum of quantitative methods used to study complex phenomena and improves the quality of research.

Keywords: extreme values, tail distribution, positional topsis socio-economic situation, territorial units

References

1. Hellwig, Z.: Zastosowania metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę wykwalifikowanych kadr. Statistical Review. 4, 307–327 (1968)
2. Hwang, C.L.; Yoon K.: Multiple attribute decision making: methods and applications. Springer-Verlag, New York (1981)
3. Łuczak, A., Just, M.: Sustainable development of territorial units: MCDM approach with optimal tail selection. Ecol. Model. 457(1), 109674 (2021)

Aleksandra Łuczak
Poznań University of Life Sciences, Faculty of Economics, Poznań, Poland,
e-mail: aleksandra.luczak@up.poznan.pl

Małgorzata Just
Poznań University of Life Sciences, Faculty of Economics, Poznań, Poland,
e-mail: ma\T1\lgorzata.just@up.poznan.pl

Depth-based Two-sample Testing

Felix Gnettner, Claudia Kirch, and Alicia Nieto-Reyes

Depth functions provide measures of the deepness of a point with respect to a given set of observations. This non-parametric concept can be applied in spaces of any dimension and entails a center-outward ordering for the given data. A two-sample test has been previously proposed that is based on depth-ranks and offers opportunities for further investigations: Observing that the corresponding test statistic $\mathcal{LS}(X,Y)$ is not symmetric with respect to the two samples X and Y, the power can be greatly increased if $\mathcal{LS}(X,Y)$ and $\mathcal{LS}(Y,X)$ are jointly considered. Within the last years, depths with respect to functional data have been established that we combine with this procedure to obtain new non-parametric two-sample tests for functional data. We investigate the asymptotic behaviour of this modified test procedure for several classes of depths including depths for functional data.

Keywords: two-sample test, depth, non-parametric

References

1. Liu, R.Y., Singh, K.: A Quality Index Based on Data Depth and Multivariate Rank Tests. Journal of the American Statistical Association. **88:421**, 252–260 (1993)

Felix Gnettner
Otto-von-Guericke-Universität Magdeburg, Institut für Mathematische Stochastik,
e-mail: felix.gnettner@ovgu.de

Claudia Kirch
Otto-von-Guericke-Universität Magdeburg, Institut für Mathematische Stochastik,
e-mail: claudia.kirch@ovgu.de

Alicia Nieto-Reyes
Universidad de Cantabria, Departamento de Matemáticas, Estadística y Computación,
e-mail: alicia.nieto@unican.es

The Control of False Discovery Rate for Functional Data

Niels Lundtorp Olsen, Alessia Pini, and Simone Vantini

In functional data analysis (FDA), the object of statistical methods are functions, which are typically modeled as random elements of a Hilbert space. In this framework inference is particularly challenging, since it deals with elements of infinite dimensional spaces. A popular topic in FDA is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of multiple comparisons problem. Local inferential techniques are either based on simultaneous confidence bands, or on the definition of a p-value function, which provides a p-value at each point of the domain, guaranteeing a control of a quantity related with the error rate on the whole domain. In this work we focus on this second line, and in particular on the control of the false discovery rate (FDR), which is the expected proportion of false discoveries (rejected null hypotheses) among all discoveries, and was first introduced in the seminal paper by Benjamini and Hochberg [1]. We define FDR in the setting of functional data defined on a manifold domain. We then introduce the functional Benjamini-Hochberg (fBH) procedure: a procedure able to control the previously defined functional FDR. Finally, the fBH procedure is applied to the analysis of daily temperatures on Earth. All details about the fBH procedure are reported in [2].

Keywords: Benjamini-Hochberg procedure, multiple comparisons, null hypothesis testing, local inference

References

1. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B, **57** (1) 289–300 (1995) doi: 10.1111/j.2517-6161.1995.tb02031.x
2. Lundtorp Olsen, N., Pini, A., Vantini, S.: False discovery rate for functional data. Test (2021)

Niels Lundtorp Olsen
DTU - Technical University of Denmark, e-mail: nalo@dtu.dk,

Alessia Pini
Università Cattolica del Sacro Cuore, e-mail: alessia.pini@unicatt.it

Simone Vantini
Politecnico di Milano, e-mail: simone.vantini@polimi.it

Functional Random Forest for Biomedical Signals Classification and Interpretative Tools

Fabrizio Maturo and Rosanna Verde

This study deals with tree-based techniques and functional data analysis (FDA) [1] for supervised classification of curves representing high-dimensional biomedical data recorded over time. Recently [2] proposed Functional Classification Trees (FCTs) and Functional Random Forest (FRF) [3] using b-spline representation and the Functional Principal Components Decomposition (FPCD) as possible basis transformation to obtain features from curves for training the classifiers. In our proposal, an original contribution is also given by new interpretative tools of the functional classification rules in the functional framework. Applications on ECG data have shown the effectiveness of the proposed functional classifiers in terms of accuracy and their usefulness in terms of interpretability.

Keywords: functional data analysis, supervised classification, functional random forest

References

1. Ramsay J, Silverman B. Functional Data Analysis, 2nd edn. New York: Springer (2005)
2. Maturo, F., Verde, R.: Pooling random forest and functional data analysis for biomedical signals supervised classification: theory and application to electrocardiogram data. Statistics in Medicine, 1–29 (2022)
3. Breiman L. Random Forests. Machine Learning **45(1)** 5–32 (2001)

F. Maturo, R. Verde
DMF, University of Campania, Caserta, Italy,
e-mail: {fabrizio.maturo,rosanna.verde}@unicampania.it

Correlation-based Iterative Clustering Methods for Time Course Data

Michelle Carey, Shuang Wu, Guojun Gan, and Hulin Wu

Many pragmatic clustering methods have been developed to group data vectors or objects into clusters so that the objects in one cluster are very similar and objects in different clusters are distinct based on some similarity measure. The availability of time course data has motivated researchers to develop methods, such as mixture and mixed-effects modelling approaches, that incorporate the temporal information contained in the shape of the trajectory of the data. However, there is still a need for the development of time-course clustering methods that can adequately deal with inhomogeneous clusters (some clusters are quite large and others are quite small).

We propose two such methods, hierarchical clustering (IHC) and iterative pairwise-correlation clustering (IPC). We evaluate and compare the proposed methods to the Markov Cluster Algorithm (MCL) and the generalised mixed-effects model (GMM) using simulation studies and an application to a time course gene expression data set from a study containing human subjects who were challenged by a live influenza virus. We identify four types of temporal gene response modules to influenza infection in humans, i.e., single-gene modules (SGM), small-size modules (SSM), medium-size modules (MSM) and large-size modules (LSM).

Keywords: functional data analysis, inhomogeneous clusters, high-dimensional data analysis

Michelle Carey
University College Dublin, Dublin, Ireland, e-mail: michelle.carey@ucd.ie

Shuang Wu
University of Rochester, New York, USA

Guojun Gan
University of Connecticut, USA

Hulin Wu
University of Texas Health Science Center School of Public Health at Houston, Houston, USA

Depth-based Classifiers for Partially Observed Functional Data

Antonio Elías, Raúl Jiménez, Anna Maria Paganoni, and Laura M. Sangalli

Partially observed functional data are frequently encountered in applications and are the object of an increasing interest by the literature. We here address the problem of classification in the context of partially observed functional data based on Depth-to-Depth classifiers [1]. To do that, we propose an integrated functional depth for partially observed functional data [2], dealing with the very challenging case where partial observability can occur systematically on any observation of the functional dataset. In particular, differently from many techniques for partially observed functional data, we do not request that some functional datum is fully observed, nor we require that a common domain exist, where all of the functional data are recorded. Because of this, our proposal can also be used in those frequent situations where reconstructions methods and other techniques for partially observed functional data are inapplicable. By means of simulation studies, we demonstrate the very good performances of the proposed depth on finite samples. We illustrate our proposal with a classification problem with data obtained from medical imaging.

Keywords: classification, partially observed functional data, depth-to-depth

References

1. Cuesta-Albertos, J. A., Febrero-Bande, M., and Oviedo de la Fuente, M.: The DD^G-classifier in the functional setting. TEST, **26**, 119–142 (2017)
2. Elías, A., Jiménez, R., Paganoni, A. M. and Sangalli, L.: Integrated depths for partially observed functional data. J. Comput. Graph. Stat.. (2022)

Antonio Elías
OASYS Group, Dept. of Applied Mathematics, Universidad de Málaga, Málaga, Spain,
e-mail: aelias@uma.es

Raúl Jiménez
Dept. of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: rjjimene@est-econ.uc3m.es

Anna M. Paganoni
MOX Laboratory for Modeling and Scientific Computing, Dept. di Matematica, Politecnico di Milano, Milano, Italy, e-mail: anna.paganoni@polimi.it

Laura M. Sangalli
MOX Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy, e-mail: laura.sangalli@polimi.it

Using Clustering and Machine Learning Methods to Provide Intelligent Grocery Shopping Recommendations

Nail Chabane, Mohamed Achraf Bouaoune, Reda Amir Sofiane Tighilt, Bogdan Mazoure, Nadia Tahiri, and Vladimir Makarenkov

Nowadays, grocery lists make part of shopping habits of many customers. With the popularity of e-commerce and plethora of products and promotions available on online stores, it can become increasingly difficult for customers to identify products that both satisfy their needs and represent the best deals overall. In this paper, we present a grocery recommender system based on the use of traditional machine learning methods aiming at assisting customers with creation of their grocery lists on the MyGroceryTour platform which displays weekly grocery deals in Canada. Our recommender system relies on the individual user purchase histories, as well as the available products' and stores' features, to constitute intelligent weekly grocery lists. The use of clustering prior to supervised machine learning methods allowed us to identify customers profiles and reduce the choice of potential products of interest for each customer, thus improving the prediction results. The highest average F-score of 0.499 for the considered dataset of 826 Canadian customers was obtained using the Random Forest prediction model which was compared to the Decision Tree, Gradient Boosting Tree, XGBoost, Logistic Regression, Catboost, Support Vector Machine and Naive Bayes models in our study.

Keywords: clustering, dimensionality reduction, grocery shopping recommendation, intelligent shopping list, machine learning, recommender systems

Nail Chabane · Mohamed Achraf Bouaoune · Reda Amir Sofiane Tighilt · Bogdan Mazoure
Université du Québec à Montreal, 405 Rue Sainte-Catherine Est, Montreal
e-mail: {chabane.nail_amine,bouaoune.mohamed_achraf}@courrier.uqam.ca;tighilt.reda@courrier.uqam.ca;bogdan.mazoure@mail.mcgill.ca

Nadia Tahiri
University of Sherbrooke, 2500 Bd de l'Université, Sherbrooke
e-mail: Nadia.Tahiri@USherbrooke.ca

Vladimir Makarenkov
Université du Québec à Montreal, 405 Rue Sainte-Catherine Est, Montreal
e-mail: makarenkov.vladimir@uqam.ca

Typology of Motivation Factors for Employees in the Banking Sector: Multivariate Data Analysis

Áurea Sousa, Osvaldo Silva, M. Graça Batista, Sara Cabral, and Helena Bacelar-Nicolau

The main purpose of this work is to know the perceptions of bank employees on the main motivational factors in the organizational context. Data analysis was performed based on Categorical Principal Component Analysis (CatPCA) and some agglomerative hierarchical clustering algorithms from *VL* parametrical family, applied to the items that aim to assess the aspects most valued by bankers. The CatPCA allowed to extract four principal components which explain almost 70% of the total data variance. The dendrograms provided by the hierarchical clustering algorithms over the same data, exhibit four main branches, which are associated with different main motivational factors. Moreover, CatPCA and clustering results show an important correspondence concerning the main motivations in this sector.

Keywords: leadership, welfare, motivational factors, catpca, cluster analysis .

References

1. Bacelar-Nicolau: The affinity coefficient. In: Bock, H.-H. and Diday, E. (eds.) Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organization, pp. 160-165. SpringerVerlag, Berlin (2000) doi: 10.1007/978-3-642-57155-8
2. Lerman, I.C.: Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series: Advanced Information and Knowledge Processing. Springer-Verlag, Boston (2016)

Áurea Sousa
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, 9500-321, Portugal
e-mail: aurea.st.sousa@uac.pt

Osvaldo Silva
Universidade dos Açores and CICSNOVA.UAc, Rua da Mãe de Deus, Portugal
e-mail: osvaldo.dl.silva@uac.pt

M. Graça Batista
Universidade dos Açores and CEEAplA, Rua da Mãe de Deus, Portugal
e-mail: maria.gc.batista@uac.pt

Sara Cabral
Universidade dos Açores, Rua da Mãe de Deus, Portugal, e-mail: sara_crc@hotmail.com

Helena Bacelar-Nicolau
Universidade de Lisboa (UL) Faculdade de Psicologia and Institute of Environmental Health (ISAMB/FM-UL), Portugal, e-mail: hbacelar@psicologia.ulisboa.pt

Industry Sector Detection in Legal Articles Using Transformer-based Deep Learning

Hui Yang, Stella Hadjiantoni, Yunfei Long, Rūta Petraitytė, and Berthold Lausen

Industry analysis, which identifies multiple industry sectors hidden in massive legal texts, could benefit greatly to business activities of legal professions such as improving customer services by detecting overall industry trends across legal topics. However, manual industry labeling on enormous legal information will be expensive and time-consuming. This research investigated an AI-powered approach to automatically recognise industry sectors using transformer-based deep learning [1] which had shown advantages on a set of Natural Language Processing (NLP) tasks recent years. In this study, a dataset consisting of over 1,700 annotated legal articles was curated for the identification of six industry sectors. Two main research questions will be answered by the outcome of our work: (1) Like other NLP tasks that mostly focus on the analysis at word token or sentence level, do transformer models also perform well on the full-text articles? (2) Compared with large-scale general domain data, could transformer models work effectively on a small legal-specific domain dataset? Our experimental results showed that the transformer-based predictive models achieved the F1 scores ranged between 0.73 and 0.83 on the detection of the six industry sectors. When the word sequence length was increased to 1,024 or 2,048 words, the performance got slight worse compared with that of the relatively short word sequences (e.g., 512). The research results suggested that transformer models are sensitive to the data size, text structure, and domain area to some extent.

Keywords: industry sector detection, text classification, transformer models

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019 Conference, pp. 4171-86. Minnesota (2019)

Hui Yang, Rūta Petraitytė
Mondaq Ltd, UK, e-mail: hui@mondaq.com;ruta.petra@mondaq.com

Stella Hadjiantoni, Berthold Lausen
Mathematical Sciences, University of Essex, UK,
e-mail: stella.hadjiantoni@essex.ac.uk;blausen@essex.ac.uk

Yunfei Long
Computer Science and Electr. Engineering, University of Essex, UK,
e-mail: yl20051@essex.ac.uk

User Segmentation Based on Online Behavioural Data via Ensemble Predictions and Clustering

Stella Hadjiantoni, Hui Yang,Yunfei Long, Ruta Petraityte, and Berthold Lausen

We use unsupervised clustering and supervised ensemble machine learning to identify segments of users defined by behavioural data. Hierarchical clustering is used as an explorative method and to identify users with similar behavioural patterns based on their online activity data (aka click data from mondaq.com). We assess estimated clusters by parametric bootstrap evaluation [1]. Stable clusters are used as additional features in ensemble prediction of win-loss probabilities for potential clients. We improve the interpretability of the machine learning model by ensembles of optimal trees [2, 3]. Our approach is compared with several machine learning models as random forest, neural networks and logistic regression.

Keywords: bootstrap assessment of cluster stability, ensemble of optimal trees, win-loss probabilities of potential clients

References

1. Lausen, B., Degens, P.O.: Evaluation of the reconstruction of phylogenies with DNA-DNA-hybridization data. In: Bock, H.-H. (ed.) Classification and Related Methods of Data Analysis: Proceedings of the First Conference of the International Federation of Classification Societies (IFCS), North-Holland, Amsterdam (1988)
2. Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., Lausen, B.: Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification **14**, 97–116 (2020)
3. Khan,Z.,Gul,N.,Faiz,N.,Gul,A.,Adler,W.,Lausen,B.:Optimal trees selection for classification via out-of-bag assessment and sub-bagging. IEEE Access **9**, 28591–28607 (2021)

Stella Hadjiantoni,
Mathematical Sciences, University of Essex, UK, e-mail: stella.hadjiantoni@essex.ac.uk

Hui Yang
Mathematical Sciences, University of Essex and Mondaq Ltd, UK, e-mail: hui@mondaq.com

Yunfei Long,
Computer Science and Electr. Engineering, University of Essex, UK ,
e-mail: yl20051@essex.ac.uk

Ruta Petraityte,
Mondaq Ltd, UK, e-mail: ruta.petra@mondaq.com

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

Attitudes Toward Statistics in the 3rd Cycle of Basic Education in Portugal

Adelaide Freitas, Ana Julieta Morais, Pedro Sá Couto, and Anabela Rocha

Statistics learning during the first school years is essential to provide all the citizens with Statistical Literacy that allows them to correctly read statistical information. While attitudes towards Statistics is a topic widely investigated at university level, it is scarce for students in the 3rd cycle of basic education (12-15 years-old). The well known questionnaire SATS-36© (Survey of Attitudes Toward Statistics, version Pos) [1] was culturally adapted to Portuguese (European) language and for students of that school cycle. Based on a sample of 215 students from schools in the central region of Portugal, a model with four factors (*Affect*, *Interest*, *Value*, and *Effort*) is proposed for the adapted scale. Randomly partitioning the data into a training sample and a test sample, the following analyses were performed to define the model: (i) Exploratory Factorial Analysis on training samples, showing invariance property of the 4-dimensional model; and (ii) Confirmatory Factorial Analysis on test samples, showing adequacy of the 4-factorial structure. The structure of the proposed model is compared with (a few) other models proposed in other countries.

Keywords: sats-36©, exploratory factorial analysis, confirmatory factorial analysis

Acknowledgements: Work partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA, University of Aveiro) through FCT (Fundação para a Ciência e a Tecnologia), reference UIDB/04106/2020.

References

1. Schau, C.: Students' attitudes: The "other" important outcome in statistics education. Ann. Mat. Pura. Appl. Proceedings of Joint Statistical Meetings. San Francisco, 3673–3681 (2003)

Adelaide Freitas and Pedro Sá Couto
CIDMA & Department of Mathematics, University of Aveiro, Campus Universitário de Santiago 3810-193 Aveiro, Portugal, e-mail: adelaide@ua.pt,p.sa.couto@ua.pt

Ana Julieta Morais
Escola Secundária Henriques Nogueira, Torres Vedras, Portugal, e-mail: ajmorais@ua.pt

Anabela Rocha
Higher Institute for Accountancy and Administration of Aveiro University, Rua Associação Humanitária dos Bombeiros Voluntários de Aveiro, 3810-902 Aveiro, Portugal,
e-mail: anabela.rocha@ua.pt

Predictors of Quantitative Skills in Degree Schemes at University

Alex Partner, Adi Lausen, Alexei Vernitski, Chris Saker, and Berthold Lausen

In the United Kingdom, students are required to study mathematics up until the age of 16. After this age it ceases to become compulsory, despite students remaining in education until the age of 18. This means that only 20% of students on UK degree schemes have studied mathematics between the ages of 16-18 [1]. Comparing this figure with over 50% uptake in comparable countries, the UK falls short in terms of maths skills in Higher Education and industry [2].
In this paper, we will discuss the findings of a two-wave study that we conducted at a UK Higher Education institution with first-year undergraduate students. We conducted the first wave of the study at the start of the university term to understand the effect mathematical literacy has on their maths and statistics performance. Further, we investigated the extent maths anxiety, personality traits and metacognition impact on their performance accuracy. Results showed that post-16 mathematics, low mathematics anxiety, low conscientiousness and low extraversion were associated with better maths and statistics performance at the start of the university term. Only higher agreeableness (working with others) was associated with higher improvement of maths and statistics performance after one term.

Keywords: personality traits, mathematics anxiety, confidence judgements

References

1. Smith, A.: Report of Professor Sir Adrian Smith's review of post-16 mathematics. London: DfE (2017)
2. Hodgen, J., Pepper, D., Sturman, L., Ruddock, D.: Is the UK an outlier? An international comparison of upper secondary mathematics education. The Nuffield Foundation (2010)

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

Adi Lausen
Department of Mathematical Sciences, University of Essex, e-mail: a.lausen@essex.ac.uk

Alexei Vernitski
Department of Mathematical Sciences, University of Essex, e-mail: asvern@essex.ac.uk

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

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

Using Excel and R for Teaching Statistics and Data Analysis

W.H. Moolman

The presentation is about the use of computer software in teaching Statistics and Data Analysis with the purpose of enhancing students learning experience and preparing them better for a career. The software that comes to mind in this regard are Excel (part of MS Office) and R (available free of charge).

The effects of Covid-19 resulted in the teaching and assessment at many universities being moved from classroom to online. With teaching and assessment online the emphasis is more on understanding concepts, computing results and interpreting computer output. In such a teaching environment the use of computer software like Excel and R is essential.

Problem: How and when to use Excel (spreadsheet based) and R (computing code based) in teaching?

Methodology: This depends very much on who are being taught and the goals to be achieved by using the software.

Results: Depending on the type of teaching, Excel and R can be used separately or to complement each other.

Implications: Excel and R can be used at every level of teaching from demonstrating simple concepts to complicated Data Mining and bootstrapping calculations.

All these issues and more will be discussed in the presentation.

Keywords: learning, spreadsheet, computing

References

1. Dell'Omodarme, M. and Valle, G.: Teaching Statistics with Excel and R. Available at https://doi.org/10.48550/arXiv.physics/0601083 (2006)
2. Duller, C., Kepler, J.: Teaching Statistics with Excel A Big Challenge for Students and Lecturers. Australian Journal of Statistics, **37**(2), 195-206 (2008)
3. Hyndman, R.: Why R is better than Excel for teaching Statistics. Part of a conversation on the Australian and New Zealand R mailing list (2010). Available at https://robjhyndman.com/hyndsight/rvsexcel/

Wessel H. Moolman
Walter Sisulu University, Mthatha, South Africa, e-mail: moolman.henri@gmail.com

Students' Assessment Through a IRT and Archetypal Analysis Joint Strategy

Lucio Palazzo and Francesco Palumbo

Item response theory [1, IRT] measures latent traits from one or more sets of manifest variables, namely items, by defining the relations between the observed variables (e.g., item responses to a test) and the latent variables. Three of the five higher education items that refer to student's abilities, as defined by the Dublin descriptors, are considered in this proposal: knowledge, application, judgment. IRT models assume that students belong to homogeneous groups concerning these abilities. Mixture IRT models [2, mixIRT] assume that the observed population is composed of latent subpopulations with class-specific quantitative parameters, representing a practical approach to finding groups by aggregating the units with respect to group's average abilities. However, assessors generally want to discover "extreme" groups of students: the most skilled, but especially those profiles that have peculiar deficits for one or more learning abilities, to define a recommendation system helping the student to fill the gaps. Archetypical analysis (AA) represents an effective data partitioning alternative to the clustering approaches around the means. The archetypes are observed or unobserved extreme points lying on the convex-hull, minimizing the sum of the squared distances from all points. The algorithm computes a membership vector for each unit with respect to each archetype. This proposal integrates the mixIRT model with the probabilistic archetypal analysis [3, PAA], presenting a hybrid estimation algorithm which iteratively computes the latent variables and the units' memberships to a set of k archetypes, where k is assumed to be known.

Keywords: item response theory, clustering, learning analytics

References

1. Hambleton, R. K., Swaminathan, H.: Item response theory, principles and applications. Springer Science & Business Media (1985)
2. Mislevy, Robert J., Verhelst, N.: Modeling item responses when different subjects employ different solution strategies. Psychometrika **55.2**, 195-215 (1990)
3. Seth, S., Eugster, M. J. A.: Probabilistic archetypal analysis. Mach Learn **102.1**, 85-113 (2016)

Lucio Palazzo
Department of Political Sciences, University of Naples Federico II, L. Rodinò 22 road - 80138 Napoli, Italy, e-mail: lucio.palazzo@unina.it

Francesco Palumbo
Department of Political Sciences, University of Naples Federico II, L. Rodinò 22 road - 80138 Napoli, Italy, e-mail: francesco.palumbo@unina.it

Kernel-based Hierarchical Structural Component Models for Pathway Analysis

Suhyun Hwangbo, Sungyoung Lee, Seungyeoun Lee, Heungsun Hwang, Inyoung Kim, and Taesung Park

Pathway analyses have led to more insight into the underlying biological functions related to the phenotype of interest in various types of omics data. Pathway-based statistical approaches have been actively developed, but most of them do not consider correlations among pathways. Because it is well known that there are quite a few biomarkers that overlap between pathways, these approaches may provide misleading results. In addition, most pathway-based approaches tend to assume that biomarkers within a pathway have linear associations with the phenotype of interest, even though the relationships are more complex.

To model complex effects including nonlinear effects, we propose a new approach, Hierarchical structural CoMponent analysis using Kernel (HisCoM-Kernel). The proposed method models nonlinear associations between biomarkers and phenotype by extending the kernel machine regression and analyzes entire pathways simultaneously by using the biomarker-pathway hierarchical structure. HisCoM-Kernel is a flexible model that can be applied to various omics data. It was successfully applied to three omics datasets generated by different technologies. Our simulation studies showed that HisCoM-Kernel provided higher statistical power than other existing pathway-based methods in all datasets. The application of HisCoM-Kernel to three types of omics dataset showed its superior performance compared to existing methods in identifying more biologically meaningful pathways, including those reported in previous studies.

Keywords: hierarchical structure, kernel, omics data, pathway

Suhyun Hwangbo
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea, e-mail: suhyun8695@gmail.com

Sungyoung Lee
Center for Precision Medicine, Seoul National University Hospital, Seoul 03080, Korea

Seungyeon Lee
Department of Mathematics and Statistics, Sejong University, Sejong 05006, Korea

Heungsun Hwang
Department of Psychology, McGill University, Montreal, QC H3A 1B1, Canada

Inyoung Kim
Department of Statistics, Virginia Tech., Blacksburg, Virginia 24060, U.S.A

Taesung Park
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea
Department of Statistics, Seoul National University, Seoul 151-747, Korea

Bayesian Inference for the Generation Interval of COVID-19 in Busan, Korea

Jayoeng Paek, Ilsu Choi, Kyeongah Nah, and Yongkuk Kim

In epidemiological dynamics, there are important parameters to provide knowledge of disease transmission, such as reproduction number, generation interval, serial interval and incubation time. Among them, generation interval, which is the transmission time difference between infector and infectee, is key parameter to estimate how quickly the disease spread out. However, it is hard to calculate generation time because the time when a person is infected is considerably uncertain. In this study, we estimate generation interval of COVID-19 in Busan, South Korea according to the emergence of new variants. Markov chain Monte Carlo (MCMC) method is used to estimate generation interval. In a simulation, distributions of generation interval and incubation period are assumed to follow gamma distribution. As a results, we provide quantiles of generation interval and pre-symptomatic transmission rates for periods with a newly predominant mutation.

Keywords: generation interval, mcmc

References

1. Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., Hens, N.: Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance **25**, 2000257 (2020)
2. Hart, W. S., Maini, P. K., Thompson, R. N.: High infectiousness immediately before COVID-19 symptom onset highlights the importance of continued contact tracing. Elife **10**, e65534 (2021)

Jayeong Paek · Ilsu Choi
Department of Mathematics and Statistics, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju, 61186, Korea, e-mail: jyp.stat@gmail.com;ichoi@jnu.ac.kr

Kyeongah Nah
Busan center for Medical Mathematics, National Institute for Mathematical Sciences, 187, Gudeok-ro, Seo-gu, Busan, 49241, Korea, e-mail: knah@nims.re.kr

Yongkuk Kim
Department of Mathematics, Kyungpook National University, 80, Daehak-ro, Daegu, 41566, Korea, e-mail: yongkuk@@knu.ac.kr

Fitting an Accelerated Failure Time Model with Time-dependent Covariates via Nonparametric Gaussian Scale Mixtures

Ju-Young Park, Byungtae Seo, and Sangwook Kang

An accelerated failure time (AFT) model is a popular regression model in survival analysis. It models the relationship between the failure time and a set of covariates via a log link with an addition of a random error. The model can be either parametric or semiparametric depending on the degree of specification of the error distribution. The covariates are usually assumed to be fixed - 'time independent'. In many biomedical studies, however, 'time-dependent' covariates are frequently observed and Cox and Oakes [2] proposed an AFT model with time-dependent covariates.
In this work, we consider a semiparametric time-dependent AFT model. We assume that the distribution of the baseline failure time as an infinite scale mixture of Gaussian densities. Thus, this model is highly flexible compared to that assumes a one-component parametric density. We consider a maximum likelihood estimation and propose an algorithm based on the constrain newton method [2] for estimating model parameters and mixing distributions. The proposed methods are investigated via simulation studies to assess the finite sample properties. The proposed methods are illustrated with a real data set.

Keywords: time dependent covariates, nonparametric gaussian-scale mixture, constrain newton method, survival analysis

References

1. Cox, D.R, and D. Oakes.: Analysis of Survival Data. Chapman & Hall, London (1984)
2. Wang, Y.: On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. Journal of the Royal Statistical Society: Series B. **69.2**, 185–198 (2007)

Ju-young Park
Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea, e-mail: jystat@yonsei.ac.kr

Byungtae Seo
Sungkyunkwan University, 25-2 Seonggyungwan-ro, Jongno-gu, Seoul, Korea
e-mail: seobt@skku.edu

Sangwook Kang
Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea, e-mail: kanggi1@yonsei.ac.kr

Comparison of Survival Prediction Models for Pancreatic Cancer: Cox Model vs. Machine Learning Models

Hyunsuk Kim, Taesung Park, and Seungyeoun Lee

A survival prediction model has been recently developed to evaluate the prognosis of nonmetastatic resected pancreatic ductal adenocarcinoma (PDAC) based on a Cox model using two nationwide database: Surveillance, Epidemiology and End Results (SEER) and Korea Tumor Registry System-Biliary Pancreas (KOTUS-BP). In this study, we applied the two machine learning methods such as random survival forests (RSF) and support vector machines (SVMs) for survival analysis, and compared the prediction performance with the Cox model, RSF and SVMs using SEER and KOTUS-BP datasets. For the model development and evaluation, three different schemes were conducted. First, we utilized data from SEER for model development and used data from KOTUS-BP for external evaluation. Secondly, these two datasets were swapped by taking data from KOTUS-BP for model development and data from SEER for external evaluation. Finally, we mixed these two datasets half and half and utilized the mixed datasets as either a model development or a validation. We used 9,624 patients from SEER and 3,281 patients from KOTUS-BP for constructing a prediction model and only seven covariates including age, sex, histologic differentiation, adjuvant treatment, resection margin status, AJCC 8th T-stage and N-stage were utilized due to the difference between sets of covariates in two datasets. Comparing the three schemes for constructing survival prediction models, the performance of Cox model, RSF and SVM is better when using mixed dataset rather than when using unmixed dataset. When using mixed dataset, the C-index, 1-year, 2-year, and 3-year time-dependent AUCs for the Cox model were 0.644, 0.698, 0.680, and 0.687, respectively. Similarly, the C-index, 1-year, 2-year, and 3-year time-dependent AUCs for RSF were 0.634, 0.682, 0.668, and 0.678, respectively while those for SVM were 0.623, 0.685, 0.635, and 0.626, respectively. In conclusion, the Cox model performs slightly better than the two machine learning methods such as RSF and SVM. This is probably because only seven clinical variables were available for constructing the prediction model.

Keywords: cox model, random survival forests, support vector machines

Hyunsuk Kim, Department of Statistics, University of California, Berkeley, USA,
e-mail: hyskim7@berkeley.edu
Taesung Park
Department of Statistics, Seoul National University, Republic of Korea,
e-mail: tspark@snu.ac.kr
Seungyeoun Lee
Department of Mathematics and Statistics, Sejong University, Republic of Korea,
e-mail: leesy@sejong.ac.kr

Clustering High-dimensional Microbiome Data

Sanjeena Dang (Subedi) and Wangshu Tu

The human microbiome plays an important role in human health and disease status. Next-generation sequencing technologies allow for quantifying the composition of the human microbiome. Clustering these microbiome data can provide valuable information by identifying underlying patterns across samples. Here, we develop a family of logistic normal multinomial factor analyzers (LNM-FA) by incorporating a factor analyzer structure. The family of models is suitable for high-dimensional microbiome data as the number of parameters in LNM-FA can be greatly reduced by assuming that the underlying latent factors is small. Parameter estimation is done using a computationally efficient variant of the alternating expectation conditional maximization algorithm that utilizes variational Gaussian approximations. The proposed method is illustrated using simulated and real datasets.

Keywords: cluster analysis, microbiome data, model-based clustering, high-dimensional data, mixture model.

Sanjeena Dang (Subedi)
School of Mathematics and Statistics, Carleton University, 1125 Colonel By Dr, Ottawa, ON K1S 5B6, e-mail: Sanjeena.Dang@carleton.ca

Wangshu Tu
School of Mathematics and Statistics, Carleton University, 1125 Colonel By Dr, Ottawa, ON K1S 5B6, e-mail: WangshuTu@cunet.carleton.ca

Clustering Adolescent Female Physical Activity Levels with an Infinite Mixture Model on Random Effects

Amy LaLonde, Tanzy Love, Deborah Rohm Young, and Tongtong Wu

Physical activity trajectories from the Trial of Activity in Adolescent Girls (TAAG) capture the various exercise habits over female adolescence. Previous analyses of this longitudinal data from the University of Maryland field site, examined the effect of various individual-, social-, and environmental-level factors impacting the change in physical activity levels over 14 to 23 years of age. We aimed to understand the differences in physical activity levels after controlling for these factors. Using a Bayesian linear mixed model incorporating a model-based clustering procedure for random deviations that does not specify the number of groups *a priori*, we find that physical activity levels are starkly different for about 5% of the study sample. These young girls are exercising on average 23 more minutes per day.

Keywords: bayesian methodology, markov chain monte carlo, mixture model, reversible jump, split-merge procedures

Amy LaLonde
University of Rochester, NY, e-mail: amylalonde2@gmail.com

Tanzy Love
University of Rochester, NY, e-mail: tanzy_love@urmc.rochester.edu

Deborah Rohm Young
University of Maryland, MD, e-mail: dryoung@umd.edu

Tongtong Wu
University of Rochester, NY, e-mail: tongtong_wu@urmc.rochester.edu

Modeling Three-way RNA Sequencing Data Using Data Transformations and Matrix-variate Gaussian Mixture Models

Theresa Scharl and Bettina Grün

RNA sequencing of time-course experiments leads to three-way count data where the dimensions are the genes, the time points and the biological units. Clustering of RNA-seq data allows to detect groups of co-expressed genes over time. After standardization, the counts of individual genes across time points and biological units constitute compositional data. Rau and Maugis [1] propose an approach for analyzing two-way RNA-seq data where they only have genes and time points as dimensions or the biological units are flattened out. For two-way data, they investigate the use of data transformations in conjunction with Gaussian mixture models. In this work we want to extend their approach to three-way data and investigate suitable data transformations for three-way data before clustering the data using matrix-variate Gaussian mixture models. Finite mixtures of matrix-variate distributions are implemented in the R package MatTransMix [2]. Using a matrix-variate Gaussian mixture model already represents a more parsimonious model formulation than using a Gaussian mixture model after flattening out the biological units. Additional parsimonity may be gained by assuming that different sets of parameters are identical across clusters, thus allowing also for an easier interpretation of the fitted model. The proposed three-way clustering approach will be applied to RNA-seq data from E. coli bioproduction processes and also compared to the two-way approach after flattening out the biological units.

Keywords: model-based clustering, three-way data, rna sequencing

References

1. Rau, A., Maugis-Rabusseau, C.: Transformation and model choice for RNA-seq co-expression analysis. Brief. Bioinformatics **19**, 425–436 (2018)
2. Zhu, X., Sarkar, S., Melnykov, V.: MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling. J Classif **39**, 147–170 (2022)

Theresa Scharl
Institute of Statistics, University of Natural Resources and Life Sciences, Peter-Jordan-Strasse 82, 1190 Vienna, Austria, e-mail: theresa.scharl@boku.ac.at

Bettina Grün
Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria, e-mail: bettina.gruen@wu.ac.at

Some Issues in Robust Clustering

Christian Hennig

Some key issues in robust clustering are discussed with focus on Gaussian mixture model based clustering, namely the formal definition of outliers, ambiguity between groups of outliers and clusters, the interaction between robust clustering and the estimation of the number of clusters, the essential dependence of (not only) robust clustering on tuning decisions, and shortcomings of existing measurements of cluster stability when it comes to outliers.

Keywords: gaussian mixture model, trimming, noise component, number of clusters, user tuning, cluster stability

References

1. García-Escudero, L. A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., Hennig, C.: Robustness and Outliers. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds.) Handbook of Cluster Analysis, pp. 653-678. Chapman & Hall/CRC, Boca Raton FL (2016)
2. Hennig, C.: Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. J. Multivariate Anal. **99**, 1154–1176 (2008)
3. Hennig, C., Coretto, P.: An adequacy approach for deciding the number of clusters for OTRIMLE robust Gaussian mixture-based clustering. Aust. N. Z. J. Stat. (2021) doi: 10.1111/anzs.12338

Christian Hennig
Dipartimento di Scienze Statistiche "Paolo Fortunati", University of Bologna, Via delle Belle Arti 41, 40126 Bologna, Italy, e-mail: christian.hennig@unibo.it

Assessing Common Principal Directions

David Rodríguez Vítores and Carlos Matrán Bea

In this paper we address the problem of comparing the principal axes of a covariance matrix with other given axes. The point of view adopted is based on the problem of optimal transport in families of location and shape, which gives rise to a very simple relation between the variances of the corresponding components in both bases. Our analysis includes the asymptotic behavior of the statistic involved, and the comparison of the method with other existing proposals in the literature.

Keywords: wasserstein distance, common principal directions, multivariate analysis, spectral functions.

References

1. Flury, B.: Common principal components and related multivariate models . Wiley, New York (1988)
2. Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: A fixed-point approach to barycenters in Wasserstein space. J. Math. Anal. Appl. **441**, 744–762 (2016)
3. Lewis, A.S., Sendov, H.S.: Twice differentiable spectral functions. SIAM J. Matrix Anal. Appl. **Vol 23, No. 2**, 368–386 (2001)

David Rodríguez Vítores
IMUVA, Universidad de Valladolid, e-mail: david.rodriguez.vitores@alumnos.uva.es

Carlos Matrán Bea
IMUVA, Universidad de Valladolid, e-mail: carlos.matran@uva.es

Robustness and Initialization Issues in Subspace Clustering

Luis A. García-Escudero and Agustín Mayo-Iscar

Observations are often assumed to cluster around lower-dimensional affine linear subspaces. In fact, this is one of the most frequently applied approaches when dealing with high or moderately high dimensional clustering problems. There are several subspace clustering approaches in the literature that attempt to find such clusters and their associated optimal underlying subspaces.

The detrimental effect that even a few outliers can have on cluster analysis, sometimes affecting even the correct determination of clusters, is well known. Robust subspace clustering methods try to discover those linear subspaces while avoiding the effect of outlying values. Detecting anomalies in the data can be an interesting problem in itself, as well as taking advantage of the subspace clustering structure to "reconstruct" the data prior to the data contamination process.

Some robustified subspace clustering methods, that follow from the application of trimming principles, will be reviewed. A proportion α of entire cases were proposed to be trimmed in [1] and a proportion α of individual cells were trimmed in [2]. These approaches provide good robustness but require the specification of a trimming rate α. A proposal will be presented to determine α based on the data.

The initialization of the iterative algorithms used to implement these trimming procedures is one of the most critical aspects for the good performance of these algorithms. Useful initialization strategies will also be provided.

Keywords: subspace clustering, robustness, high-dimensions

References

1. García-Escudero, L.A., Gordaliza, A., San Martin, R., Van Aelst, S., Zamar, R.: Robust linear clustering. J. R. Stat. Soc. Ser. B **71**, 301–318 (2009)
2. García-Escudero, L. A., Rivera-García, D., Mayo-Iscar, A., Ortega, J.: Cluster analysis with cellwise trimming and applications for the robust clustering of curves. Inf. Sci. **573**, 100–124 (2021)

Luis A. García-Escudero
Dpto. Estadística e I.O. and IMUVA, University of Valladolid, Valladolid 47011, Spain
e-mail: lagarcia@uva.es

Agustín Mayo-Iscar
Dpto. Estadística e I.O. and IMUVA, University of Valladolid, Valladolid 47011, Spain
e-mail: agustin.mayo.iscar@uva.es

A Likelihood Ratio Test for Choosing Input Parameters in Robust Model Based Clustering

Luis A. García-Escudero, Agustín Mayo-Iscar, Gianluca Morelli, and Marco Riani

In the last twenty five years robust several proposals for maximum likelihood estimation based on trimming and constraints have been developed. For these procedures, consistency results have been obtained and their robustness has been justified.

There remains an open issue, when applying estimators based on the joint application of trimming and constraints, related to choosing the number of clusters, the level of trimming and the strength of the constraints imposed on the components' scatter matrices. Some exploratory tools are available to help users make these decisions using so-called "ctlcurves".

A new parametric bootstrap-based likelihood ratio test procedure has been developed to identify combinations of input parameters associated to the most interesting clustering solutions. The statistical properties of this proposal and empirical evidence on its performance when applied to artificial and real data, including contaminating observations, will be presented.

Keywords: model based clustering, trimming, constrained estimation

Luis A. García-Escudero
Dpto. Estadística e I.O. and IMUVA, University of Valladolid, Valladolid 47011, Spain,
e-mail: lagarcia@uva.es

Agustín Mayo-Iscar
Dpto. Estadística e I.O. and IMUVA, University of Valladolid, Valladolid 47011, Spain,
e-mail: agustin.mayo.iscar@uva.es

Gianluca Morelli
Department of Economics and Management and Interdepartmental Centre of Robust Statistics, University of Parma, Italy, e-mail: gianluca.morelli@unipr.it

Marco Riani
Department of Economics and Management and Interdepartmental Centre of Robust Statistics, University of Parma, Italy, e-mail: mriani@unipr.it

Data Clustering and Representation Learning Based on Networked Data

Lazhar Labiod and Mohamed Nadif

To deal simultaneously with both, the attributed network embedding and clustering, we propose a new model exploiting both content and structure information. The proposed model relies on the approximation of the relaxed continuous embedding solution by the true discrete clustering. Thereby, we show that incorporating an embedding representation provides simpler and easier interpretable solutions. Experiment results demonstrate that the proposed algorithm performs better, in terms of clustering, than the state-of-art algorithms, including deep learning methods devoted to similar tasks.

Keywords: networked data, clustering, representation learning, spectral rotation

Lazhar Labiod
Université de Paris, CNRS, Centre Borelli UMR 9010, e-mail: lazhar.labiod@u-paris.fr

Mohamed Nadif
Université de Paris, CNRS, Centre Borelli UMR 9010, e-mail: mohamed.nadif@u-paris.fr

Exploratory Graph Analysis for Configural Invariance Assessment of a Test

Alex Cucco, Lara Fontanella, Sara Fontanella, and Nicola Pronello

Self-report survey instruments are frequently used to investigate differences between groups of respondents, such as citizens of different nations in cross-country comparative analyses. In this context, a main methodological problem pertain to the configural invariance of the measurement instrument, which holds if the latent structure has the same pattern across different groups across the groups. In this work, to address this issue, we adopt an exploratory approach rooted in the framework of graph theory. Specifically, considering a multi-group comparative analysis and measurement instruments consisting of ordered categorical indicators, we discuss the use of exploratory graph analysis to assess the instrument configural invariance. In this framework, networks are used to represent latent constructs, and the covariance between observable indicators is explained in terms of a pattern of causal interactions between the items. Hence, we assume that if the measurement instrument functions invariantly across the groups, the group specific correlation-based networks will be characterised by a similar structure. The network structures are estimated through a Bayesian approach with sparse inducing priors and network embedding will be used to investigate the structure similarity. Through a simulation study we demonstrate that the proposed method is able to identify the differences. Finally, the proposed approach is applied to test the configural invariance of the Democracy Scale adopted in the European Social Survey.

Keywords: psychometric networks, bayesian sparse modelling, dimensionality reduction

Alex Cucco
National Heart and Lung Institute, Imperial College London, London, UK,
e-mail: a.cucco20@imperial.ac.uk

Lara Fontanella
Department of Legal and Social Sciences, University of Chieti-Pescara, Chieti, Italy,
e-mail: lara.fontanella@unich.it

Sara Fontanella
National Heart and Lung Institute, Imperial College London, London, UK,
e-mail: s.fontanella@imperial.ac.uk

Nicola Pronello
Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Chieti, Italy, e-mail: nicola.pronello@studenti.unich.it

An Extension of Edge Reduction for Large Networks

Pedro Campos

In Network Science, edge reduction in graphs has been studied by several authors for many years [2] [1]. The applications are varied, from telecommunications, to Internet of Things (IoT), and fraud detection. In fraud detection, the objective is to simplify the structure of networks to better identify money laundering patterns. In most literature, edge reduction techniques are based on network structure and link weights. In this work we use node attributes and edge attributes to reduce the structure of large graphs, where the edges and nodes are characterized by a large amount of features. For the reduction task we use an adaptation of Supervised PCA, an algorithm that uses a subset of features based on their association with the outcome ([4]),but we extend the edge reduction by using a double reduction both at the levels of nodes and edges using a double Supervised PCA (2X-SPCA). An illustrative application of the method is made with a variant of PaySim, a Synthetic Financial Dataset For Fraud Detection ([3]), containing more than 6 million of transactions (the edges) between more than 2 million users (the nodes). An outcome variable - isFraud – is used, assigning the edge to a transaction made by the fraudulent agents inside the simulation.

Keywords: edge reduction, double supervised PCA, fraud detection

References

1. Papageorgiou, A., Cheng, B., and Kovacs, E., "Real-time data reduction at the network edge of Internet-of-Things systems," 2015 11th International Conference on Network and Service Management (CNSM), 2015, pp. 284-291, doi: 10.1109/CNSM.2015.7367373.
2. Hambrusch, S.E., Lim, H. (1999). Minimizing the Diameter in Tree Networks Under Edge Reductions. Parallel Process. Lett., 9, 361-371.
3. Lopez-Rojas, E. A., Elmir, A., and Axelsson, S., "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016
4. Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006). Prediction by Supervised Principal Components. Journal of the American Statistical Association, 101(473), 119–137. http://www.jstor.org/stable/30047444
5. amburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**, 321–354 (1995)

Pedro Campos
Faculty of Economics, University of Porto, and LIAAD INESC TEC, R Dr. Roberto Frias, Porto, e-mail: pcampos@fep.up.pt

Patterns of Cooperation for Polish Authors of Research Publications in Economics, Business and Medicine Areas

Urszula Cieraszewska, Paweł Lula, Magdalena Talaga, and Marcela Zembura

The analysis of the publication activity not only has great cognitive importance, but it should be treated as an important tool supporting scientific activity management. The authors' attention has been directed to this aspect of scientific productivity research, which concerns the analysis of main features of the network of cooperation among authors and the identification of the most frequent patterns of cooperation [1, 2].

The authors are planning to:

1. identify the type of networks in the context of random networks (Erdős - Rényi - Gilbert model) and small-world networks (Watts–Strogatz model),
2. carry out an analysis of the structure of author teams and determine their size, diversification of institutions and countries represented by authors and lastingness of teams,
3. study the relationship between essential features of authors' teams and main measures reflecting publication significance expressed by journal prestige and the number of citations of a given paper,
4. compare networks of authors and patterns of cooperation for economics, business and medicine area.

The analysis will be performed using data from Scopus database describing research publications of Polish authors working in the field of economics, business and medicine.

Keywords: scientific productivity, patterns of cooperation, network models

References

1. Cook, D. J., Holder, L. B. (eds.): Mining Graph Data. Wiley-Interscience (2010)
2. Pieńkosz, K., Wojciechowski, J.: Grafy i sieci. Wydawnictwo Naukowe PWN, Warszawa (2013)

Urszula Cieraszewska, Paweł Lula, Magdalena Talaga
Cracow University of Economics, Poland,
e-mail: cieraszu@uek.krakow.pl, lulap@uek.krakow.pl, talagam@uek.krakow.pl

Marcela Zembura
Medical University of Silesia, Poland, e-mail: marcela.zembura@gmail.com

A Topological Clustering of Individuals

Rafik Abdesselam

The clustering of objects-individuals is one of the most widely used approaches to exploring multidimensional data. The two common unsupervised clustering strategies are Hierarchical Ascending Clustering (HAC) and k-means partitioning used to identify groups of similar objects in a dataset to divide it into homogeneous groups. The proposed Topological Clustering of Individuals, or TCI, studies a homogeneous set of individuals-rows of a data table, based on the notion of neighborhood graphs; the columns-variables are more-or-less correlated or linked according to whether the variable is of a quantitative or qualitative type. It enables topological analysis of the clustering of individual variables which can be quantitative, qualitative or a mixture of the two. It first analyzes the correlations or associations observed between the variables in the topological context of principal component analysis (PCA) or multiple correspondence analysis (MCA), depending on the type of variable, then classifies individuals into homogeneous groups relative to the structure of the variables considered. The proposed TCI method is presented and illustrated here using a simple real dataset with quantitative variables; however, it can also be applied with qualitative or mixed variables.

Keywords: hierarchical clustering, proximity measure, neighborhood graph, adjacency matrix, multivariate data analysis

References

1. Abdesselam, R.: A Topological Principal Component Analysis. International Journal of Data Science and Analysis. Vol.7, Issue 2, 20–31 (2021)
2. Batagelj, V., Bren, M.: Comparing resemblance measures. Journal of classification, 12, 73–90 (1995)
3. Lesot, M. J., Rifqi, M. and Benhadda, H.: Similarity measures for binary and numerical data: a survey. In: IJKESDP, 1, 1, 63-84 (2009)
4. Panagopoulos, D.: Topological data analysis and clustering. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning, (2022)
5. Zighed, D., Abdesselam, R., and Hadgu, A.: Topological comparisons of proximity measures. In: Tan et al. (Eds). 16th PAKDD 2012 Conference, pp. 379–391. Springer, (2012)

Rafik Abdesselam
University of Lyon, Lyon 2, ERIC - COACTIS Laboratories, Department of Economics and Management, 69365 Lyon, France, e-mail: rafik.abdesselam@univ-lyon2.fr

Modeling a Most Specific Generalization in Domain Taxonomies

Zhirayr Hayrapetyan, Boris Mirkin, Susana Nascimento, trevor Fenner, and Dmitry Frotov

We define a most specific generalization of a fuzzy set of topics assigned to leaves of the rooted tree of a domain taxonomy. This generalization lifts the set to its "head subject" node in the higher ranks of the taxonomy tree. The head subject is supposed to "tightly" cover the query set, possibly involving some errors referred to as "gaps" and "offshoots". We develop a method to globally maximize either the parsimony or the likelihood of a scenario involving gains and losses of the general concept manifested in a fuzzy cluster of leaf nodes of the taxonomy. Supplemented with fuzzy c-means clustering, this allows us to obtain meaningful generalizations for fuzzy thematic clusters of Data Science topics using several dozen thousand abstracts from issues of relevant research journals published from 2000 on.

Keywords: generalization, maximum parsimony, maximum likelihood, fuzzy thematic cluster, research tendencies

References

1. The 2012 ACM Computing Classification System — Association for Computing Machinery.—2012.— URL: https://www.acm.org/publications/class-2012.
2. E. Chernyak, B. Mirkin. Refining a taxonomy by using annotated suffix trees and Wikipedia resources // Annals of Data Science. Vol. 2, no. 1, 61-82, 2015.
3. D. Frolov, S. Nascimento, T.I. Fenner, B. Mirkin. Parsimonious generalization of fuzzy thematic sets in taxonomies applied to the analysis of tendencies of research in Data Science// Information Sciences, 512, pp. 595-615, 2020.

Zhirayr Hayrapetyan
National Research University Higher School of Economics, Moscow, Russia
e-mail: zhayrapetyan@ithse.ru

Boris Mirkin
National Research University Higher School of Economics, Moscow, Russia, Birkbeck, University of London, London, UK, e-mail: bmirkin@hse.ru

A Proposal for Formalization and Definition of Anomalies in Dynamical Systems

Jan Michael Spoor, Jens Weber, and Jivka Ovtcharova

Although many scientists strongly focus on anomaly detection in different applications and domains, there currently exists no universally accepted definition of anomalies and outliers [1]. Using an approach based on control theory and dynamical systems, as well as a definition for anomalies as described by philosophy of science [2], the authors propose a generalized framework viewing anomalies as key drivers of progress for a better understanding of the dynamical systems around us. By mathematically defining anomalies and delimiting deviations within expectations from completely unforeseen instances, this paper aims to be a contribution to set up a universally accepted definition of anomalies and outliers.

Keywords: anomaly detection, outlier analysis, dynamical systems

References

1. Hodge, V.J.; Austin, J.A.: Survey of Outlier Detection Methodologies. Artif Intell Rev 22, pp. 85-126 (2004)
2. Spoor, J.M.; Weber, J.; Ovtcharova, J.: A Definition of Anomalies, Measurements and Predictions in Dynamical Engineering Systems for Streamlined Novelty Detection. Accepted for the 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul (2022)

Jan Michael Spoor
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jan.spoor@kit.edu

Jens Weber
Team Digital Factory Sindelfingen, Mercedes-Benz Group AG, Sindelfingen, Germany
e-mail: jens.je.weber@mercedes-benz.com

Jivka Ovtcharova
Institut für Informationsmanagement im Ingenieurwesen (IMI), Karlsruhe Institute of Technology, Karlsruhe, Germany, e-mail: jivka.ovtcharova@kit.edu

Unsupervised Classification of Categorical Time Series Through Innovative Distances

Ángel López-Oriona, José A. Vilar, and Pierpaolo D'Urso

In this paper, two novel distances for nominal time series are introduced. Both of them are based on features describing the serial dependence patterns between each pair of categories. The first dissimilarity employs the so-called association measures, whereas the second computes correlation quantities between indicator processes whose uniqueness is guaranteed from standard stationary conditions. The metrics are used to construct crisp algorithms for clustering categorical series. The approaches are able to group series generated from similar underlying stochastic processes, achieve accurate results with series coming from a broad range of models and are computationally efficient. An extensive simulation study shows that the devised clustering algorithms outperform several alternative procedures proposed in the literature. Specifically, they achieve better results than approaches based on maximum likelihhod estimation, which take advantage of knowing the real underlying procedures. Both innovative dissimilarities could be useful for practitioners in the field of time series clustering.

Keywords: categorical time series, clustering, association measures, indicator processes

Ángel López-Oriona, José A. Vilar
Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, Spain, e-mails: oriona38@hotmail.com, jose.vilarf@udc.es

Pierpaolo D'Urso
Department of Social Sciences and Economics, Sapienza University of Rome, Italy
e-mail: pierpaolo.durso@uniroma1.it

Detecting Differences in Italian Regional Health Services During Two Covid-19 Waves

Lucio Palazzo and Riccardo Ievoli

During the first two waves of Covid-19 pandemic, territorial healthcare systems have been severely stressed in many countries. The availability (and complexity) of data requires proper comparisons for understanding differences in performance of health services. We apply a three-steps approach to compare the performance of Italian healthcare system at territorial level (NUTS 2 regions), considering daily time series regarding both intensive care units and ordinary hospitalizations of Covid-19 patients. Changes between the two waves at a regional level emerge from the main results, allowing to map the pressure on territorial health services.

Keywords: regional healthcare, time series, multidimensional scaling, cluster analysis, trimmed k-means

Lucio Palazzo
Department of Political Sciences, University of Naples Federico II, L. Rodinò 22 road - 80138 Napoli, Italy, e-mail: lucio.palazzo@unina.it

Riccardo Ievoli
Department of Chemical, Pharmaceutical and Agricultural Sciences, University of Ferrara, via Luigi Borsari 46 - 44121 Ferrara, Italy, e-mail: riccardo.ievoli@unife.it

The Clustering Performance of a Weighted Combined Distance Between Time Series

Margarida G. M. S. Cardoso, Ana Alexandra Martins, and João Lagarto

Recently, [1], we proposed a new dissimilarity measure between time series - COMB, a uniform convex combination of four (normalized) distance measures: Euclidean; Pearson correlation based; Periodogram based; and a distance between estimated autocorrelation structures. In this work, we propose a method to determine the weights of the convex combination of distances in COMB: it relies on the concordance of clusterings obtained by each individual distance measure and COMB derived clustering. A weighted COMB measure is thus obtained, WCOMB. We then test the clustering performance of WCOMB vs. COMB by conducting an experimental analysis on all the time series datasets of the UCR archive. We evaluate the concordance between the clusters obtained using K-Medoids and the original classes (using adjusted Rand index) as well as the cohesion-separation of the clusters (using the Silhouette index). In addition, we consider a clustering application - with data from the Portuguese Transmission System Operator, on time series of electricity consumption (2014 to 2019) - to compare the performance of both methods. Significant differences between the average Silhouette values of clusters obtained were found. The concordance with the original classes' structure exhibits similar performance in both approaches. We conclude that, for unsupervised leaning, it can be worthwhile to invest on deriving specific weights for the distances integrating COMB.

Keywords: clustering, distance measures, time series

References

1. Cardoso, M. G. M. S., Martins, A. A.: The performance of a combined distance between time series. In: Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R. and de Carvalho, M. (eds.) Recent Developments in Statistics and Data Science - Proceedings of the XXV Congress of the Portuguese Statistical Society. Springer.(to be published)

Margarida G. M. S. Cardoso
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research UnitName of Institute, Avenida das Forças Armadas, 1649-026 Lisbon, Portugal e-mail: margarida.cardoso@iscte-iul.pt

Ana Alexandra Martins
CIMOSM, Instituto Superior de Engenharia de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007, Lisbon, Portugal e-mail: ana.martins@isel.pt

João Lagarto
Instituto Superior de Engenharia de Lisboa and INESC-ID, Rua Conselheiro Emídio Navarro, 1, 1959-007, Lisbon, Portugal e-mail: jlagarto@deea.isel.ipl.pt

Dimensionality Reduction and Multivariate Time Series Classification

Veronne Yepmo, Angeline Plaud, and Engelbert Mephu Nguifo

In this work we tackle the problem of dimensionality reduction when classifying multivariate time series (MTS). Multivariate time series classification is a challenging task, especially as sparsity in raw data, computational runtime and dependency among dimensions increase the difficulty to deal with such complex data.

In a recent work, a novel subspace model named EMMV (Ensemble de M-histogrammes Multi-Vues) [1] that combines M-histograms and multi-view learning together with an ensemble learning technique to handle the MTS classification task was reported. The aforementioned model has shown good results when compared to state of the art MTS classification methods. Before performing the classification itself, EMMV reduces the dimension of the multivariate time series using correlation analysis, and uses after that a random selection of the views. In this work, we explore two more alternatives to the dimensionality reduction method used in EMMV, the goal being to check the efficiency of randomness on EMMV. The first technique named Temporal Laplacian Eigenmaps [2] comes from manifold learning and the second one named Fractal Redundancy Elimination [3] comes from the fractal theory. Both are nonlinear dimensionality reduction algorithms in contrast to correlation analysis which is linear, meaning that the first cited are able to eliminate more correlations than the latter.

We then conduct several experiments on available MTS benchmarks in order to compare the different techniques, and discuss the obtained results.

Keywords: multivariate time series, dimensionality reduction, classification

Acknowledgements This work was partially supported by LabEx IMobS3 and IMI2-H2020 Project NeuroDeRisk.

References

1. Plaud, A., Mephu Nguifo, E. and Charreyron, J.: Classification des séries temporelles multivariées par l'usage de Mgrams. French Machine Learning conf. (CAP), 2019, July, Toulouse https://hal.archives-ouvertes.fr/hal-02162093.
2. Lewandowski, M., et al.: Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. 20th IEEE ICPR, 2010. p. 161-164.
3. Oliveira, J. and Cordeiro, R.: Unsupervised dimensionality reduction for very large datasets: Are we going to the right direction ? Knowledge-Based Systems. 2020, vol. **196**, p. 105777

Veronne Yepmo, Angeline Plaud, and Engelbert MEphu Nguifo
Univ. Clermont Auvergne, Clermont Auvergne INP, LIMOS, ISIMA, Clermont-Ferrand, France,
e-mail: engelbert.mephu_nguifo@uca.fr

A Review on Official Survey Item Classification for Mixed-Mode Effects Adjustment

Afshin Ashofteh and Pedro Campos

The COVID-19 pandemic has had a direct impact on the development, production, and dissemination of official statistics. This situation led National Statistics Institutes (NSIs) to make methodological and practical choices for survey collection without the need for the direct contact of interviewing staff (i.e. remote survey data collection). Mixing telephone interviews (CATI) and computer-assisted web interviewing (CAWI) with direct contact of interviewing constitute a new way for data collection at the time COVID-19 crisis. This paper presents a literature review to summarize the role of statistical classification and design weights to control coverage errors and non-response bias in mixed-mode questionnaire design. We identified 289 research articles with a computerized search over two databases, Scopus and Web of Science. It was found that, although employing mixed-mode surveys could be considered as a substitution of traditional face-to-face interviews (CAPI), proper statistical classification of survey items and responders is important to control the nonresponse rates and coverage error risk.

Keywords: mixed-mode official surveys, item classification, weighting methods, clustering, measurement error

References

1. Ashofteh, A., and Bravo, J. M.: A study on the quality of novel coronavirus (COVID-19) official datasets. Stat. J. IAOS, vol. 36, no. 2, pp. 291–301, (2020). doi: 10.3233/SJI-200674
2. Ashofteh, A., and Bravo, J. M.: Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. Stat. J. IAOS, vol. 37, no. 3, pp. 771–789, (2021). doi: 10.3233/SJI-200674
3. Kim, S. and Couper,M. P.: Feasibility and Quality of a National RDD Smartphone Web Survey: Comparison With a Cell Phone CATI Survey. Soc. Sci. Comput. Rev., vol. 39, no. 6, pp. 1218–1236, (2021).

Afshin Ashofteh
Statistics Portugal (Instituto Nacional de Estatística, Departamento de Metodologia e Sistemas de Infomação) , and NOVA Information Management School (NOVA IMS) and MagIC, Universidade Nova de Lisboa, Portugal,
e-mail: afshin.ashofteh@ine.pt e-mail: aashofteh@novaims.unl.pt

Pedro Campos
Statistics Portugal (Instituto Nacional de Estatística, Departamento de Metodologia e Sistemas de Infomação), and University of Porto, Faculty of Economics. Universidade do Porto, Portugal, e-mail: pedro.campos@ine.pt

Adaptive Fuzzy Systems in Economics and Finance: Evaluating Interval Forecasts of High-frequency Data

Rosangela Ballini

The forecast of the future movement of economic and financial variables assumes a central role for the composition of portfolios, risk management, asset pricing and investment analysis; therefore, the development of prediction methodologies is of fundamental importance. With the recent and rapid growth in the availability of financial information, especially at intraday frequencies, approaches to forecasting interval time series have gained prominence in the literature, since they comprise the construction of more informative forecasts, capable of capturing the fluctuations of an asset, index or rate over the course of a transaction day, as opposed to techniques based on one-off anticipations. In general, forecast models have some practical limitations, such as linear structure, nonconsideration of market uncertainties, restrictive hypotheses about the data, need for a large number of observations to estimate parameters, and inadequacy for the natural treatment of interval data. To address such limitations, fuzzy modeling has been proposed for forecasting interval time series. Adaptive fuzzy models are nonlinear, are able to update their structure and functionality according to data streams, and handle uncertainty from fuzzy sets. Thus, this approach allows the dynamic treatment of complex phenomena, as well as considering information affected by uncertainties, as is the case of financial markets. These approaches have been applied to different economies by forecasting stock prices, exchange rates, stock exchange indices and price volatility, as opposed to traditional univariate and multivariate econometric methods.

Keywords: adaptive fuzzy systems, time series forecasting, high-frequency data

Rosangela Ballini
Institute of Economics, University of Campinas, Brazil, e-mail: ballini@unicamp.br

The Usefulness of Selected Machine Learning Methods for Estimating Missing Data to Supplement Databases Used for Corporate Bankruptcy Prediction

Barbara Pawelek and Jozef Pociecha

Socio-economic consequences of corporate bankruptcies and forecasting the risk firms' bankruptcy enjoys unflagging interest among researchers and practitioners. Financial indicators that are the basis for building forecasting models very often contain data shortages. In our previous research, we supplemented the missing data primarily by median of a given variable, determined separately for bankrupt and non-bankrupt firms [2]. The aim of the paper is presentation the results of investigations on the usefulness of selected machine learning methods for estimation values of missing data (e.g. multivariate imputation method based on random forests), to supplement databases used for building and estimations corporate bankruptcy prediction models. The study included three databases designed to forecast the bankruptcy of enterprises in Poland one year, two years and three years in advance. Various mechanisms for generating data gaps were considered [1]. Selected methods were used to estimate missing data, such as: mean, median, k-nearest neighbors, classification tree, multivariate imputation based on random Forests or predictive mean matching [3], and others.

Keywords: missing data, bankruptcy prediction, machine learning

References

1. Kauermann, G., Küchenhoff, H., Heumann, C.: Statistical Foundations, Reasoning and Inference. For Science and Data Science. Springer, Cham (2021)
2. Pociecha, J., Pawelek, B., Baryla, M., Augustyn, S.: Classification Models as Tools of Bankruptcy Prediction – Polish Experience. In: Mola, F., Conversano, C., Vichi, M. (eds.) Classification, (Big) Data Analysis and Statistical Learning. Springer, Cham (2018)
3. van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. **45**(3), 1-67 (2011)

Barbara Pawelek
Cracow University of Economics, Department of Statistics, 31-510 Kraków, ul. Rakowicka 27, Poland, e-mail: barbara.pawelek@uek.krakow.pl

Jozef Pociecha
Cracow University of Economics, Department of Statistics, 31-510 Kraków, ul. Rakowicka 27, Poland, e-mail: jozef.pociecha@uek.krakow.pl

Registration of 24-hour Accelerometric Rest-activity Profiles and its Application to Human Chronotypes

Erin I. McDonnell, Vadim Zipunnikov, Jennifer A. Schrack, Jeff Goldsmith, and Julia Wrobel

By collecting data continuously over 24 hours, accelerometers and other wearable devices can provide novel insights into circadian rhythms and their relationship to human health. Existing approaches for analyzing diurnal patterns using these data, including the cosinor model and functional principal component analysis, have revealed and quantified population-level diurnal patterns, but considerable subject-level variability remained uncaptured in features such as wake/sleep times and activity intensity. This remaining informative variability could provide a better understanding of chronotypes, or behavioral manifestations of one's underlying 24-hour rhythm. Curve registration, or alignment, is a technique in functional data analysis that separates "vertical" variability in activity intensity from "horizontal" variability in time-dependent markers like wake and sleep times; this data-driven approach is well-suited to studying chronotypes using accelerometer data. We develop a parametric registration framework for 24-hour accelerometric rest-activity profiles represented as dichotomized into epoch-level states of activity or rest. Specifically, we estimate subject-specific piecewise linear time-warping functions parametrized with a small set of parameters. We apply this method to data from the Baltimore Longitudinal Study of Aging and illustrate how estimated parameters give a more flexible quantification of chronotypes compared to traditional approaches.

Keywords: functional data analysis, alignment, clustering

Erin I. McDonnell
Google, United States, e-mail: eim2117@cumc.columbia.edu

Vadim Zipunnikov
Johns Hopkins University, United States, e-mail: vzipunni@jhsph.edu

Jennifer A. Schrack
Johns Hopkins University, United States, e-mail: jschrac1@jhu.edu

Jeff Goldsmith
Columbia University Mailman School of Public Health, United States
e-mail: ajg2202@cumc.columbia.edu

Julia Wrobel
University of Colorado Denver, United States, e-mail: julia.wrobel@cuanschutz.edu

Functional Data from Wearable Devices: a Review

Nihan Acar-Denizli and Pedro Delicado

With the recent development of sensor and information technologies, it is more and more common to collect data obtained from people by sensors or wearable devices in a continuous an automatic way, to which we refer as *wearable device data*. Mobile phones have sensors and accelerometers measuring from the number of steps that we have walk daily to our instantaneous stress level. In healthcare there are wearable devices that measure the amount of oxygen in the blood, the electrical activity of the heart over time and the concentration of glucose in blood. See [2].

Since wearable device data are usually continuous, time-dependent, and has high volume, Functional Data Analysis (FDA) are suitable for them. Through an exhaustive literature revision, here we explore the possibilities that FDA offers as a generic methodology for analyzing wearable device data. In particular, we identify relevant problems in wearable data that can be approached using FDA, and we document open access datasets that can be used as benchmarks in posterior research on functional data coming from wearable devices.

As an example, in [1] data come from a clinical trial evaluating the beneficial effects of quinoa consumption on prediabetic subjects, which were monitored during 8 weeks with FreeStyle Libre (a sensor applied to the back of the upper arm of subjects that records the data on glucose concentration every 15 minutes). The glucose values corresponding to the breakfast (from minute -30' to +120') were considered the functional response in a functional regression model fitted to measure the effect on glucose curves of diet type (rich in quinoa or not) and nutrient intakes.

Keywords: accelerometer, glucose level curve, remote patient monitoring

References

1. Diaz-Rizzolo, D. A., Acar-Denizli, N., Kostov, B., Roura, E., Sisó-Almirall, A., Delicado, P., Gomis, R.: Beneficial effects of quinoa consumption on glycaemia fluctuations: a pilot study in old-age prediabetic subjects. Submitted. (2022)
2. Goldsmith, J., Liu, X., Jacobson, J. S., Rundle, A.: New insights into activity patterns in children, found using functional data analyses. Med. Sci. Sports Exerc., **48**, 1723–1729 (2016).

Nihan Acar-Denizli
Universitat Politècnica de Catalunya, Barcelona, Spain, e-mail: nihan.acar.denizli@upc.edu

Pedro Delicado
Universitat Politècnica de Catalunya, Barcelona, Spain, e-mail: pedro.delicado@upc.edu

A Wavelet-mixed Effect Landmark Model for the Effect of Potassium and Biomarkers Profiles on Survival in Heart Failure Patients

Caterina Gregorio, Giulia Barbati, and Francesca Ieva

Statistical methods to study the association between a longitudinal biomarker and the risk of death are a very relevant problem for the long-term monitoring of biomarkers. In this context, sudden crises can cause the biomarker to undergo very abrupt changes. Although these oscillations are typically short-term, they often contain relevant prognostic information. We propose a method that couples a linear mixed-model with a wavelet smoothing to extract both the long-term component and the short-term oscillations of the individual longitudinal biomarker profiles, and describe them as functional data.

We then use them as predictors in a landmark model to study their association with the risk of death. To illustrate the method, we use clinical application which motivated our work, i.e. the monitoring of potassium and related biomarkers in Heart Failure patients. The dataset consists of real-world data coming from the integration of Administrative Health Records and Outpatient and Inpatient Clinic E-chart from Trieste (Italy).

Our method not only allows us to identify the short-term oscillations, but also reveals their prognostic role, according to their duration, demonstrating the importance of including them in the modeling. Compared to other state of the art methods (e.g., landmark analyses and joint models), our proposal archives higher predictive performances. Our analysis has also an important clinical implications, since it allows us to derive a dynamic score that can be used in clinical practice to assess the risk related to an observed patient's potassium trajectory and then tune the actual drug therapy she/he has to undergo.

Keywords: mixed-effect models, landmark survival analysis, time-dependent covariates, functional data, heart failure

Caterina Gregorio
MOX - Department of Mathematics, Politecnico di Milano, 20133 Milano (IT)
e-mail: caterina.gregorio@polimi.it

Giulia Barbati
Biostatistic Unit, Department of Medical Sciences, University of Trieste, 34100 Trieste (IT)
e-mail: gbarbati@units.it

Francesca Ieva
MOX - Department of Mathematics, Politecnico di Milano, 20133 Milano (IT),
CHDS - Center for Health Data Science, Human Technopole, 20156 Milano (IT)
e-mail: francesca.ieva@polimi.it

True Sparsity Approaches in Classification via Conic Optimization

Immanuel M. Bomze and Bo Peng

Pursuing sparsity is an important issue in all classification tasks, in particular in view of the nowadays increasing popular move towards explainable machine learning. Here we address this quest by linking the exact sparsity term/zero norm

$$\|\mathbf{x}\|_0 = \text{number of nonzero } x_i \text{ 's}$$

to copositive optimization. We present a novel, purely continuous model, which avoids any branching or use of large constants in implementation. The resulting model is a (nonconvex) quadratic optimization problem with complementarity constraints. We show that the copositive formulation is exact under mild conditions involving only the constraints, not the (classifying criterion) objective, and discuss strong duality to ensure tight bounds. The covered problem class includes sparse least-squares regression under linear constraints as well. Numerical comparisons between our method and other approximations are reported from the perspective of criterion value.

Keywords: sparse classifier, constrained least squares, conic optimization

Immanuel M. Bomze
ISOR/VCOR, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria,
e-mail: immanuel.bomze@univie.ac.at

Bo Peng
VGSCO, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria,
e-mail: bo.peng@univie.ac.at

Creating Homogeneous Sectors: Criteria and Applications of Sectorization

Cristina Lopes, Maria Margarida Lima, Elif Göksu Öztürk, Ana Maria Rodrigues, Ana Catarina Nunes, Cristina Oliveira, José Soeiro Ferreira, and Pedro Filipe Rocha

Sectorization is the process of grouping a set of previously defined basic units (points or small areas) into a fixed number of sectors. Sectorization is also known in the literature as *districting* or *territory design* [1], and is usually performed to optimize one or more criteria regarding the geographic characteristics of the territory and the planning purposes of sectors. The most common criteria are equilibrium, compactness and contiguity, which can be measured in many ways [2].

Sectorization is similar to clustering but with a different motivation. Both aggregate smaller units into groups. But, while clustering strives for inner similarity of data, sectorization aims at outer homogeneity [1]. In clustering, groups should be very different from each other, and similar points are classified in the same cluster. In sectorization, groups should be very similar to each other, and therefore very different points can be grouped in the same sector.

We classify sectorization problems into four types: basic sectorization, sectorization with service centers, resectorization, and dynamic sectorization. A Decision Support System for Sectorization, D3S, is being developed to deal with these four types of problems. Multi-objective genetic algorithms were implemented in D3S using Python, and a user-friendly web interface was developed using Django. Several applications can be solved with D3S, such as political districting, sales territory design, delivery service zones, and assignment of fire stations and health services to the population.

Keywords: sectorization, clustering, decision support system, optimization

References

1. Kalcsics, J., Nickel, S., Schroeder,M.: Towards a unified territorial design approach - Applications, algorithms and GIS integration. TOP **13**(1), 1–56 (2005)
2. Rodrigues, A.M., Ferreira, J.S.: Measures in Sectorization Problems. In: Póvoa, A., de Miranda, J. (eds.) Operations Research and Big Data **15**. Springer, Cham. (2015)

Cristina Lopes, Maria Margarida Lima, Cristina Teles and Ana Maria Rodrigues
ISCAP, Polytechnic of Porto, Portugal, e-mail: cristinalopes@iscap.ipp.pt

Ana Maria Rodrigues, Elif Göksu Öztürk, José Soeiro Ferreira and Pedro F. Rocha
INESCTEC - Technology and Science, Porto, Portugal e-mail: ana.m.rodrigues@inesctec.pt

Ana Catarina Nunes
ISCTE - University Institute of Lisbon, and CMAFcIO - Faculty of Sciences, University of Lisbon, Lisbon, Portugal e-mail: catarina.nunes@iscte-iul.pt

MARGOT: a Maximum MARGin Optimal Classification Tree

Federico D'Onofrio, Marta Monaci, Giorgio Grani, and Laura Palagi

In recent years there has been a growing attention to machine learning models which are able to give explanatory insights on the decisions made by the algorithm. Thanks to their interpretability, decision trees have been intensively studied for classification tasks, and, due to the remarkable advances in mixed-integer programming (MIP), various approaches have been proposed to formulate the Optimal Classification Tree (OCT) problem as a MIP model starting from the seminal paper [2]. We present a novel MIQP formulation for binary classification using OCT and exploiting the generalization capabilities of Support Vector Machines. The maximum MARGin Optimal Classification Tree (MARGOT) selects at each node of the decision tree a maximum margin separating hyperplane using an ℓ_2-norm linear SVM (see e.g. [1] and references therein). The resulting model combines such multivariate hyperplanes minimizing the global misclassification error. The model can also include feature selection constraints, following e.g. [3], which allows to define a hierarchy on the subsets of features which mostly affect the outcome. MARGOT has been tested on non-linearly separable synthetic datasets in a 2-features space in order to provide a graphical representation of the optimal hyperplanes. Finally, MARGOT has been tested on benchmark datasets from the UCI repository.

Keywords: support vector machines, optimal decision trees, mixed integer quadratic programming

References

1. Piccialli, V. and Sciandrone, M.: Nonlinear optimization and support vector machines. Ann Oper Res (2022).
2. Bertsimas, D. and Dunn, J.: Optimal classification trees. Machine Learning, **7**, 1039–1082 (2017).
3. Labbé, M., and Martínez-Merino, L. I. and Rodríguez-Chía, A. M.: Mixed integer linear programming for feature selection in support vector machine. Discrete Applied Mathematics **261**, 276–304 (2019).

Federico D'Onofrio · Marta Monaci · Laura Palagi
Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, Via Ariosto 25, Rome 00185, Italy,
e-mail: \{federico.donofrio,marta.monaci,laura.palagi\}@uniroma1.it

Giorgio Grani
Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, Rome 00185, Italy, e-mail: g.grani@uniroma1.it

Multivariate Mapping of Soil Organic Carbon and Nitrogen

Stephan van der Westhuizen, David P. Hofmeyr, and Gerard B.M. Heuvelink

Soil maps, which can be effectively produced with statistical models in digital soil mapping (DSM), contain vital information on the spatial distribution of soil properties which are used in fields such as water- and land management and climate studies [1]. Soil maps are usually produced in a univariate manner, that is, each map is produced independently and therefore, when multiple soil properties are mapped the underlying covariance structure between these soil properties is ignored. This may lead to inconsistent soil maps, for example, organic carbon and nitrogen maps produced independently may show unrealistic carbon-nitrogen ratios. The latter is important as these ratios are used by map users to obtain information on residue decomposition and the nitrogen cycle in the soil. In the last decade the production of soil maps with machine learning models has become increasingly popular as these models are able to quantify complex non-linear relationships between a soil property and the environmental covariates. However, producing soil maps with multivariate machine learning models is still lacking and requires much investigation in DSM. In this talk we present the simultaneous mapping of soil organic carbon and nitrogen for the region consisting of Belgium, The Netherlands, Luxembourg, and Germany. The simultaneous mapping is performed with a multivariate random forest model [2], and we compare this model to that of two separate univariate random forest models.

Keywords: random forest, digital soil mapping, multivariate analysis, machine learning

References

1. McBratney, A. B., Mendonça Santos, M. L., Minasny, B.: On digital soil mapping. Geoderma. **117**, 3–52 (2003)
2. Segal, M., Xiao, Y.: Multivariate random forests. WIREs Data Min. Knowl. Discovery. **1**, 80–87 (2011)

Stephan van der Westhuizen
Stellenbosch University, Stellenbosch, South Africa, e-mail: stephanvdw@sun.ac.za

David P. Hofmeyr
Stellenbosch University, Stellenbosch, South Africa

Gerard B.M. Heuvelink
Wageningen University and Research, Wageningen, The Netherlands

Spatial Configuration of Fire Stations in Portugal

Regina Bispo, Clara Yokochi, Francisca G. Vieira, Nádia Bachir, Pedro Espadinha-Cruz, José Pedro Lopes, Alexandre Penha, Marta Belchior Lopes, Filipe J. Marques, João Paulo Rodrigues, and António Grilo

Fire stations (FS) provide a global emergency response to non-fire incidents, e.g., vehicle crashes. Urban fires are nonetheless one of the most frequent types of occurrences and sources of property damage that may lead to severe losses. This severity is linked to the characteristic higher population densities and larger building agglomerations in urban centers. In Portugal, FS are very non-uniformly spatially distributed between municipalities both in terms of number and geographical location. Since the spatial configuration of fire stations may considerably influence the effectiveness of the provided services, national and regional governments need research-based advice on how many and where to establish firefighting facilities. Hence, based on information regarding urban and rural fires and vehicle crashes occurrences, this study aimed at estimating the number of FS per municipality by fitting a Poisson regression while accounting for spatial dependence. In addition, an unsupervised machine learning clustering approach was used to assess the adequacy and efficiency of FS locations, aiming to minimize the distance or the arrival time to the incidents.

Keywords: fire, fire stations, Poisson regression, k-means

Regina Bispo, Nádia Bachir, Filipe J. Marques
NOVAMATH Center for Mathematics and Applications, Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA da Lisboa, Caparica, Portugal
e-mail: r.bispo@fct.unl.pt

Clara Yokochi, Francisca G. Vieira
NOVA School of Science and Technology, Universidade NOVA da Lisboa, Caparica, Portugal

Pedro Espadinha-Cruz, António Grilo
UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA da Lisboa, Caparica, Portugal

José Pedro Lopes
Escola Nacional de Bombeiros and Instituto Superior de Educação e Ciências, Coimbra, Portugal

Alexandre Penha
Comando Nacional de Emergência e Proteção Civil, ANEPC Autoridade Nacional de Emergência e Proteção Civil, Carnaxide, Portugal

Marta B. Lopes
NOVAMATH Center for Mathematics and Applications, NOVA Laboratory for Computer Science and Informatics NOVALINCS, NOVA School of Science and Technology, Universidade NOVA da Lisboa, Caparica, Portugal

João Paulo Rodrigues
Department of Civil Engineering, University of Coimbra, Coimbra, Portugal

Spatio-temporal Variability of Distribution and Abundance of Sardine in Portuguese Continental Coast: Environmental Effects

Daniela Silva, Raquel Menezes, Ana Moreno, Ana Teles-Machado, and Susana Garrido

Scientific tools capable of identifying the distribution patterns of species are important as they contribute to improve knowledge about biodiversity and species abundance, to make sustainable management decisions and conserve biodiversity. This study aims to estimate the spatio-temporal distribution of sardine (*Sardina pilchardus*, Walbaum 1792) in the Western Iberian waters and Gulf of Cadiz, relating the spatio-temporal variability of biomass indicator with the environmental conditions. Acoustic data was collected during Portuguese spring acoustic (PELAGO) surveys conducted by the Portuguese Institute for Sea and Atmosphere (IPMA) over a total of 19920 hauls from 2000 to 2020 (gap in 2012). Daily environmental data was obtained for the region and time of study, particularly satellite derived sea surface temperature, chlorophyll-a concentration, bathymetry, and intensity and direction of ocean currents. Species Distribution Models are investigated to relate sardine presence/absence and biomass with the environmental conditions, aiming at predicting sardine distribution in unobserved locations and for the unobserved year of 2012. The hurdle Bayesian models become suitable since they allow to incorporate the specificities of the data: complex spatio-temporal dynamics, excess of zeros, and the difference between the occurrence and biomass under occurrence processes. The hurdle model is a two-part model such that species biomass is given by the product of these two processes. In addition to considering the spatio-temporal structure, the impact of the covariates with a time lag on biomass indicator is evaluated using a kernel gaussian function. Data from the west and south Iberian coasts are studied separately due to the shape of the coast and the different oceanographic conditions.

Keywords: environmental effects, geostatistics, hurdle model, *sardina pilchardus*, species distribution model

Daniela Silva
Minho University, Braga, Portugal, e-mail: danyelasylva2@gmail.com

Raquel Menezes
Minho University, Guimarães, Portugal, e-mail: rmenezes@math.uminho.pt

Ana Moreno
Portuguese Institute for Sea and Atmosphere, Algés, Portugal, e-mail: amoreno@ipma.pt

Ana Teles-Machado
Portuguese Institute for Sea and Atmosphere, Algés, Portugal, e-mail: ana.machado@ipma.pt

Susana Garrido
Portuguese Institute for Sea and Atmosphere, Algés, Portugal, e-mail: susana.garrido@ipma.pt

Time Resolved Feature Importance of a Biopharmaceutical Purification Process Using Permutation Based Methods

Matthias Medl, Theresa Scharl, Astrid Dürauer, and Friedrich Leisch

Real-time monitoring of critical process parameters of biotechnological processes is a major step towards quality-by-design in the production of biopharmaceuticals. The emergence of novel monitoring devices has resulted in the accumulation of complex high-dimensional data. Recently, statistical models capable of predicting critical process parameters online -e.g. the product or impurity concentration- have been developed. However, to generate these models the variable space has been reduced manually based on expert knowledge. This presents a problem as (a) expert knowledge is not always available, especially for novel technologies, (b) experts might overlook important variables and (c) the importance of some variables might be unknown in general. Therefore, we propose a deep learning framework capable of predicting critical process parameters of a biopharmaceutical purification process based on the whole high-dimensional variable space (>1400 variables). To achieve this, a neural network architecture consisting of two parallel strands that are concatenated at the end has been developed. One strand consists of fully connected layers and takes standalone variables -e.g. pH, conductivity- as input, while the other strand consists of convolutional layers and utilizes whole Fourier transform infrared spectra as input. Using this method, the model itself learns, which variables contain useful information or not. By determining the variable importances with the model, (a) previously unknown correlations and patterns can be identified to gain further understanding about the underlying mechanics of the purification process and (b) more accurate models can be generated that utilize all informative variables available.

Keywords: deep learning, variable importance, permutation, bioprocess

Matthias Medl
Institute of Statistics, University of Natural Resources and Life Sciences, Peter-Jordan-Strasse 82, 1190 Vienna, Austria, e-mail: matthias.medl@boku.ac.at

Theresa Scharl
Institute of Statistics, University of Natural Resources and Life Sciences, Peter-Jordan-Strasse 82, 1190 Vienna, Austria, e-mail: theresa.scharl@boku.ac.at

Astrid Dürauer
Institute of Bioprocess Science and Engineering, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria, e-mail: astrid.duerauer@boku.ac.at

Friedrich Leisch
Institute of Statistics, University of Natural Resources and Life Sciences, Peter-Jordan-Strasse 82, 1190 Vienna, Austria, e-mail: friedrich.leisch@boku.ac.at

Off-target Predictions in CRISPR-Cas9 Gene Editing Using Machine Learning

Ali Mertcan Kose and Ozan Kocadağlı

Recently, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeated) applications widely appear in gene editing for treatment of cancer. Therefore, CRISPR Cas-9 system is a robust method for effectively editing the genome of cell. CRISPR locus is composed of DNA in genes located on cell. DNA/Target sequence occurs 23 endonucleases. Off/on- target sequence are diagnosed with matching between endonucleases and Guide RNA. The prediction of off-target mutations in CRISPR-Cas9 is a hot topic because of its relevance to gene editing research. In literature,Off/on target levels are often evaluated by CFD/MIT scores in terms of binary classification. Instead of those scores, determining more than 2 classes by the latent class analysis (LCA)in the pre-processing step helps to interpret, and classify more accurately. In this study, a benchmark dataset that consists of Human (Homo-sapiens) - (GRCh37/hg19) + SNPs: 10000 Genomes, ExaC was used. In the analysis, LCA produced three significant classes related to the off-target scores over benchmark dataset. These classes are identified as high (7.1%), middle (86.5%), low (6.4%). Afterwards, the estimated off-target scores were modeled by machine learning methods such as Xgboost, SVM, ANN and decision trees etc. where the benchmark dataset was partitioned by 10-fold cross-validation procedure. The analysis results figure out the effect of locus structures and mismatching positions on the off-target.The best model is Xgboost with accuracy(AUC=100%).

Keywords: latent class analysis, machine learning, CRISPR, off target scoring

References

1. Kang, S. H., Lee, W. jae, An, J. H., Lee, *et al.*: Prediction-based highly sensitive CRISPR off-target validation using target-specific DNA enrichment. Nat Commun. **11**(1), 1–11 (2020)
2. Leibowitz, M. L., Papathanasiou, S., Doerfler, *et al.*:Chromothripsis as an on-target consequence of CRISPR–Cas9 genome editing. Nat Genet. **53**(6), 895–905 (2021)

Ali Mertcan Kose
Mimar Sinan Fine Arts University, Silahsor Cad. No:71 MSGSU Bomonti Campus, 34380 Sisli/Istanbul, Turkey, e-mail: alimertcankose@gmail.com

Ozan Kocadağlı
Mimar Sinan Fine Arts University, Silahsor Cad. No:71 MSGSU Bomonti Campus, 34380 Sisli/Istanbul, Turkey, e-mail: ozan.kocadagli@msgsu.edu.tr

Comparison of k-mer and Alignment-based Pre-processing Approaches for Machine Learning Based Functional Annotation with 16S rRNA Data

Rafal Kulakowski, Adi Lausen, Etienne Low-Decarie, and Berthold Lausen

Over the last few decades, the continuing advancements in Next-Generation-Sequencing technologies provided new opportunities to obtain large volumes of biological sequence data from uncultured environments. Moreover, continuing efforts to store labelled RNA, DNA and protein sequences have opened opportunities to implement Machine Learning (ML) techniques and build predictive tools that can estimate key characteristics of sequenced environments. Traditionally, processing of biological sequences for comparative analysis involved implementing sequence alignment techniques. However, alignment algorithms have high computational costs that scale non-linearly with the number of sequences. Moreover, current Multiple Sequence Alignment methods produce a representation of data, where a format of each sequence is heavily dependent on other most similar sequences in the current set, making any subsequently trained predictive model unstable.
Our investigation focused on identifying scalable and effective data pre-processing techniques for the series of functional annotation tasks using 16S rRNA data [1]. To this end, we tested the use of a pairwise-alignment pre-processing technique, which we then compared to an alignment-free, k-mer based method. Additionally, we examined whether combining both techniques improves the accuracy of ML classifiers. The results of our experiments showed that the k-mer frequencies provide the most favourable set of features for these problems.

Keywords: sequence alignment, k-mer, feature engineering, functional annotation, classification

References

1. Kulakowski, R., Lausen, A, Low-Decarie, E., Lausen, B.: Classification Methods for 16S rRNA Based Functional Annotation. Archives of Data Science, Series A **4** (1), 23 (2020).

Rafal Kulakowski
Imperial College Health Partners, UK,
e-mail: rafal.kulakowski@imperialcollegehealthpartners.com

Adi Lausen
Mathematical Sciences, University of Essex, UK, e-mail: a.lausen@essex.ac.uk

Etienne Low-Decarie
Canadian Space Agency, Canada, e-mail: etienne.low-decarie@canada.ca

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

An Ultrametric Model for Clustering and Dimensionality Reduction

Giorgia Zaccaria

The study of multidimensional phenomena is currently growing with the complexity of the reality, raising the need for methodologies to explore the relationships between their many facets. Multidimensional phenomena are often explained by nested latent concepts ordered in a hierarchical, tree structure, whose characterization can differ in heterogeneous populations. In this work, a new parsimonious parameterization of the covariance matrix able to pinpoint a hierarchical structure on variables is proposed, and implemented into a Gaussian Mixture Model (GMM). The proposal is based upon the definition of an ultrametric matrix [2], which is one-to-one associated with a hierarchy of latent concepts. Its implementation into a GMM aims, on one hand, at introducing a new parsimonious GMM with a reduced number of parameters and, on the other hand, at identifying a different characterization of the phenomenon under study for each component (subpopulation) of the mixture. With respect to the existing parsimonious parameterizations of the component covariance structure, e.g., the eigen-decomposition [2] and the decomposition based on Factor Analysis [1], the ultrametric GMM works particularly well in situations where a hierarchy over variables can be identified. Nonetheless, the proposal shows good performance also when a general (non-hierarchical) covariance structure is assumed for the data. The application of the proposal to real data concerning well-being and a benchmark data set illustrates its potentials to explore multidimensional phenomena in a heterogeneous population.

Keywords: ultrametric models, parsimonious parameterization, model-based clustering, hierarchy of latent concepts

References

1. Banfield, J., Raftery, A.: Model-based Gaussian and non-Gaussian clustering. Biometrics. **49**(3), 803-–821 (1993)
2. Dellacherie, C., Martinez, S., San Martin, J.: Inverse M-matrices and ultrametric matrices. Lecture Notes in Mathematics, Springer, Switzerland (2014)
3. McLachlan, G, Peel, D.: Mixtures of factor analyzers. In: Langley, P. (ed.) Proceedings of the seventeenth international conference on machine learning, pp 599–606. Morgan Kaufmann, San Francisco (2000)

Giorgia Zaccaria
University of Rome Unitelma Sapienza, Piazza Sassari 4, 00161, Rome, Italy,
e-mail: giorgia.zaccaria@unitelmasapienza.it

Combining Latent Class Analysis and Multiple Correspondence Analysis

Alice Barth

Both Latent Class Analysis (LCA) and Multiple Correspondence Analysis (MCA) are analysis methods to find patterns in complex categorical data tables. LCA aims at representing heterogeneity in the observed data by estimating internally homogeneous groups (the latent classes). It can thus be understood as a probability-based clustering approach. MCA is a scaling method that reduces the dimensionality of a multi-way frequency table. As such, it is an extension of simple correspondence analysis (CA) for two-way frequency tables. Both CA and MCA are often used to visualize relations in a lower-dimensional space. Several possibilities of combining LCA and correspondence analysis have been discussed. Some authors use MCA as a diagnostic tool to select variables which are subsequently used in an LCA. Others propose to visualize results from an LCA using CA [1].
In this presentation, a different approach is discussed: the projection of LCA results as passive variables into a two-dimensional space created by MCA, using the example of relations between attitudes towards migration and socio-demographic characteristics in Germany (World Values Survey Round 7, 2017). First, latent structures in attitudes towards migration are estimated based on eight items such as "immigration in your country increases unemployment" with answer options "agree", "disagree" and "hard to say". Information criteria indicate a four-class-solution. Second, a "social space" is constructed via MCA. Here, socio-demographic characteristics such as gender, highest educational qualification, occupational group and size of town of residence are used. The association of latent class membership with certain socio-demographic characteristics is assessed by projecting the categories of the latent class variable (modal posterior probabilities) into the two-dimensional MCA solution. Further, possibilities of preserving information on individuals' probabilities of class membership in this approach are discussed.

Keywords: categorical data, latent class analysis, correspondence analysis

References

1. McCutcheon, A.: Correspondence Analysis Used Complimentary to Latent Class Analysis in Comparative Social Research. In: Blasius, J., Greeenacre, M. (eds.) Visualization of Categorical Data, pp. 477-488. Academic Press (1998)

Alice Barth
Institute of Sociology, Bonn University, Germany, e-mail: albarth@uni-bonn.de

Simultaneous Factorial Reduction and Clustering for Three-mode Data Sets: a Comparison

Prosper Ablordeppey, Adelaide Freitas, Maurizio Vichi, and Giorgia Zaccaria

Two iterative techniques, called T3Clus and 3Fkmeans, aimed at a simultaneous clustering of objects and a factorial dimensionality reduction of variables and occasions on three-mode data sets, as well as a combination of these two procedures were proposed in [1]. In each iteration, T3Clus (3Fkmeans) is based on a sequential application of the Tucker2 algorithm and the k-means algorithm (vice versa). We have implemented the three simultaneous methods in Python. In this work, applications on real data sets are presented to show the features of these three simultaneous methods when compared with tandem analyses which can be executed in two different ways. In a tandem analysis only one sequential application of clustering and factorial methodologies is performed.

Keywords: three-mode data, clustering, factorial dimensionality reduction

Acknowledgements This work was partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA, University of Aveiro) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), reference UIDB/04106/2020.

References

1. Vichi, M., Rocci, R., Kiers, H.: Simultaneous Component and Clustering Models for Three-way Data: Within and Between Approaches. Journal of Classification **24**, 71–98 (2007)

Prosper Ablordeppey and Adelaide Freitas
CIDMA & Department of Mathematics, University of Aveiro, Portugal,
e-mail: {pablordeppey,adelaide}@ua.pt

Maurizio Vichi and Giorgia Zaccaria
Department of Statistical Sciences, University "La Sapienza", Rome, Italy,
e-mail: maurizio.vichi@uniroma1.it,giorgia.zaccaria@uniroma1.it

Clustering Intensive Longitudinal Data Through Mixture Multilevel Vector-autoregressive Modeling

Anja Ernst, Marieke Timmerman, Feng Ji, Bertus Jeronimus, and Casper Albers

Experience sampling methodology is increasingly used in the social sciences to analyze individuals' emotions, thoughts and behaviors in everyday-life. The resulting intensive longitudinal data is often analyzed with the objective to describe the inter-individual differences that are present within it. To accommodate inter-individual differences to a greater extent than previously possible, a mixture multilevel vector-autoregressive model is proposed. This model combines a mixture model at level 2 (individual level) with a multilevel vector-autoregressive model [1] that describes the dynamic fluctuations present at level 1 (time-point level). This exploratory model identifies mixture components of individuals who exhibit similar overall means, autoregressions, and cross-regressions. Within each mixture component, multilevel coefficients allow additionally for within-component variation on these vector-autoregressive coefficients. The advantage of exploratory identifying mixture components and accounting for within-component variation is demonstrated on data from the COGITO study. This data contains samples of individuals from disparate age groups of over 100 individuals each.

Keywords: model-based clustering, time series analysis, applications in social sciences

References

1. Rovine, M. J., Walls, T. A.: Multilevel Autoregressive Modeling of Interindividual Differences in the Stability of a Process. In: Walls, T. A., Schafer, J. L. (eds.) Models for intensive longitudinal data, pp. 124–147. Oxford University Press (2006)

Anja Ernst
University of Groningen, The Netherlands, e-mail: a.f.ernst@rug.nl

Marieke Timmerman
University of Groningen, The Netherlands, e-mail: m.e.timmerman@rug.nl

Feng Ji
University of California, Berkeley, USA, e-mail: fengji@berkeley.edu

Bertus Jeronimus
University of Groningen, The Netherlands, e-mail: b.f.jeronimus@rug.nl

Casper Albers
University of Groningen, The Netherlands, e-mail: c.j.albers@rug.nl

Mispecification Tests for Hidden Markov Models Based on a New Class of Finite Mixture Models

Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni

In the context of longitudinal data, we show that a general class of hidden Markov (HM, [1]) models may be equivalent to a class of finite mixture (FM, [3]) models based on an augmented set of components and suitable constraints on the conditional response probabilities, given these components. We formulate a misspecification test for the latent structure of an HM model comparing maximum likelihood values of the two models for the same data, and when the number of possible latent state sequences is excessive, we propose a multiple version of this test including the Bonferroni correction. The procedure is simple since it is based on the output of the Expectation-Maximization estimation algorithm [2]. The properties of this testing procedure are evaluated through a simulation study. An empirical application illustrates it through data from the National Longitudinal Survey of Youth, in which we jointly consider wages and years of experience after labour force entry. We show that the proposed testing procedure may also be used as an alternative model selection criterion for the number of latent states of an HM model to those usually employed.

Keywords: expectation-maximization algorithm; likelihood ratio test; model selection; multiple testing

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Boca Raton FL: Chapman & Hall/CRC. Taylor & Francis (2013)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. **39**, 1–38, (1977)
3. McLachlan, G., Lee, S.X., Rathnayake, S.I.: Finite mixture models. Annu. Rev. Stat. Appl. **10**, 355–378 (2019)

Francesco Bartolucci and Silvia Pandolfi
Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia, Italy,
e-mail: francesco.bartolucci@unipg.it, e-mail: silvia.pandolfi@unipg.it

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via degli Arcimboldi, 8, 20126 Milano, Italy, e-mail: fulvia.pennoni@unimib.it

Natural Cubic Smoothing Splines for Latent Class Identification in Longitudinal Growth Trajectories

Katerina M. Marcoulides and Laura Trinchera

Latent growth curve modeling is regularly used to examine intra-individual changes over time, inter-individual differences in intra-individual changes over time, as well as a variety of other intra- and inter-individual disparities over time. In such models, the growth trajectories are usually modeled as linear functions. However, nonlinear patterns of change over time are quite common in social and behavioral science research. Recently Marcoulides and Khojasteh [2] proposed to use natural cubic smoothing splines to analyze non-linear longitudinal data. This approach has the main advantage of avoiding knot selection when using splines and avoids imposing overly restrictive assumptions that other non-linear modeling approaches often require. One limitation is that all the sampled individuals in a given longitudinal study are assumed to rise from a single population. Traditional growth mixture modeling methods are useful when analyzing such samples with unobserved heterogeneity, though they still assume the growth trajectories to be the same for all individuals within a latent class.[1]. During our presentation we will discuss a novel approach that uses derivatives of individual natural cubic smoothing spline functions and then, following some recent work by Marcoulides and Trinchera [3], applies a hierarchical clustering algorithm to group or cluster individuals who follow similar growth trajectory patterns without requiring to define the number of classes a priori.

Keywords: unobserved heterogeneity, SEM, latent class detection

References

1. Diallo, T. M. O., Morin, A.J.S., Lu, H.: Impact of misspecifications of the latent variance-covariance and residual matrices on the class enumeration accuracy of growth mixture models. Struct. Equ. Model. **23**, 507–531(2016)
2. Marcoulides, K.M., Khojasteh, J.: Analyzing longitudinal data using natural cubic smoothing splines. Struct. Equ. Model. **25**, 965–971(2018)
3. Marcoulides, K.M., Trinchera, L.: Detecting unobserved heterogeneity in latent growth curve models. Struct. Equ. Model. **26**, 390–401 (2018)

Katerina M. Marcoulides
University of Minnesota, Department of Psychology, Minneapolis, MN, 55455, USA,
e-mail: kmarcoul@umn.edu

Laura Trinchera
NEOMA Business School, 1 Rue du Marchal Juin, 76130 Mont-Saint-Aignan, France,
e-mail: laura.trinchera@neoma-bs.fr

Supervised Classification via Neural Networks for Replicated Point Patterns

Kateřina Pawlasová, Iva Karafiátová, and Jiří Dvořák

A spatial point pattern is a collection of points observed in a bounded region of $\mathbb{R}^d$, $d \geq 2$. Individual points represent, e.g., observed locations of cell nuclei in a tissue ($d = 2$) or centers of undesirable air bubbles in industrial materials ($d = 3$). The main goal of this paper is to show the possibility of solving the supervised classification task for point patterns via neural networks with general input space [3]. To predict the class membership for a newly observed pattern, we compute an empirical estimate of a selected functional characteristic (e. g., the pair correlation function). Then, we consider this estimated function to be a functional variable that enters the input layer of the network. A short simulation example illustrates the performance of the proposed classifier in the situation where the observed patterns are generated from two models with different spatial interactions. In addition, the proposed classifier is compared with convolutional neural networks [1] (with point patterns represented by binary images) and kernel regression. Kernel regression classifiers for point patterns have been studied in our previous work [2], and we consider them a benchmark in this setting.

Keywords: spatial point patterns, pair correlation function, supervised classification, neural networks, functional data

References

1. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
2. Koňasová, K., Dvořák, J.: Techniques from functional data analysis adaptable for spatial point patterns (2021) Available as a part of the Proceedings of the 22nd European Young Statisticians Meeting. https://www.eysm2021.panteion.gr/publications.html.Cited10Jan2022
3. Thind, B., Multani, K., Cao, J.: Deep Learning with Functional Inputs (2020) Available via arxiv. https://arxiv.org/pdf/2006.09590.pdf.Cited10Jan2022

Kateřina Pawlasová
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: pawlasova@karlin.mff.cuni.cz

Iva Karafiátová
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: karafiatova@karlin.mff.cuni.cz

Jiří Dvořák
Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Praha 2, Czech Republic, e-mail: dvorak@karlin.mff.cuni.cz

Reliability Assessment of Ancient Stone Arch Bridge Applying ANN models, Case Study: Leça Railway Bridge

Edward A. Baron, Ana Margarida Bento, José Campos e Matos, Rui Calçada, and Kenneth Gavin

In many areas of engineering, surrogate models are used to replace traditional methods of obtaining data and evaluating the performance of engineering outcomes. However, the investment in computation time using analytical models is, in some cases, impractical (a simulation could take minutes, hours, or even days). This is the case for the structural assessment field, where the construction of approximation models to predict the performance by generating a relationship between the analyzed parameters (inputs and outputs) is recommended. The main objective of this investigation is to evaluate a surrogate model into a developed data set of a bridge case study. In this sense, the data set will be generated by using a finite element model to consider structural uncertainties to compute the ultimate load-carrying capacity. Thus, based on the data set, a mathematical function will be derived to compute failure probabilities through statistical analysis [1].

Keywords: ancient bridge, neural network, probability of failure, reliability index

References

1. Baron, E. A., Galvão, N., Docevska, M., Matos, J., Markovski, G.: Application of quality control plan to existing bridges. Struct. Infrastruct. Eng., DOI: 10.1080/15732479.2021.1994618 (2021)

Edward A. Baron
University of Minho, ISISE, Department of Civil Engineering, 4800-058 Guimarães, Portugal, e-mail: id8033@alunos.uminho.pt

Ana Margarida Bento
University of Minho, ISISE, Department of Civil Engineering, 4800-058 Guimarães, Portugal, e-mail: ana.bento@civil.uminho.pt

José Campos e Matos
University of Minho, ISISE, Department of Civil Engineering, 4800-058 Guimarães, Portugal, e-mail: jmatos@civil.uminho.pt

Rui Calçada
University of Porto, CONSTRUCT, Faculty of Engineering, 4099-002, Porto, Portugal, e-mail: ruiabc@fe.up.pt

Kenneth Gavin
Delft University of Technology, Faculty of Civil Engineering and Geosciences, 2628 CD Delft, The Netherlands, e-mail: k.g.gavin@tudelft.nl

Application of Artificial Intelligence (AI) in Flood Risk Forecasting

Minh Quang Tran, Ana Margarida Bento, Elisabete Teixeira, Hélder Sousa, and José Campos e Matos

Floods are the most serious natural disasters currently, directly affecting people's daily lives and causing many serious effects. The number of annual floods is increasing with different intensity in several locations [1]. The sustainable development of people, critical infrastructure, the economy and society will be severely affected if no preventive or countermeasures are taken [2]. With the goal of minimizing damage by providing effective flood prevention and response solutions, forecasting requires high accuracy and long forecasting time. However, recently, along with the powerful development of computer science, big data and the development history of society, artificial intelligence (AI) tools have the potential to perform this challenging task more accurately and faster than traditional methods, as highlighted in Pham et al. [3]. This investigation presents the application of AI in a comprehensive flood risk forecasting methodology, with the potential to provide a useful tool for flood management and definition of mitigation measures in urban areas, as well as for assisting in catastrophic events prevention.

Keywords: artificial intelligence (AI), flood events, predictive model

References

1. The EU Floods Directive. https://ec.europa.eu/environment/water/flood_risk/.
2. Kundzewicz, Z.W., Hegger, D.L.T., Matczak, P., Driessen, P.P.J. (2018): Flood-risk reduction: Structural measures and diverse strategies. Proceedings of the National Academy of Sciences.
3. Pham, B.T., Luu, C., Tran, P., Nguyen, H.D., Le, H.V., Quoc, T., Ta, H.T., Prakash, I. (2021): Flood risk assessment using hybrid artificial intelligence models integrated with multi-criteria decision analysis in Quang Nam Province, Vietnam. Journal of Hydrology 592(11).

Minh Quang Tran
Institute for Sustainability and Innovation in Structural Engineering, University of Minho, 4800-058 Guimarães, Portugal, e-mail: minhtq@uct.vn

Ana Margarida Bento
ISISE, University of Minho, Guimarães, Portugal, e-mail: ana.bento@civil.uminho.pt

Elisabete Teixeira
ISISE, University of Minho, Guimarães, Portugal, e-mail: elisabeterodriguest@gmail.com

Hélder Sousa
ISISE, University of Minho, Guimarães, Portugal, e-mail: sousa.hms@gmail.com

José Campos e Matos
ISISE, University of Minho, Guimarães, Portugal, e-mail: jmatos@civil.uminho.pt

Logistic Regression with Sparse Common and Distinctive Covariates

Soogeun Park, Eva Ceulemans, and Katrijn Van Deun

Having large sets of predictor variables from multiple sources concerning the same individuals is becoming increasingly common in research. On top of the variable selection problem, predicting the category in which the observations belong to using such data gives rise to an additional challenge of identifying the processes at play underneath the predictors. These processes are of particular interest in the setting of multi-source data because they can either be associated individually with a single data source or jointly with multiple sources. Although many methods have addressed the classification problem in high dimensionality, the additional challenge of distinguishing such underlying predictor processes from multi-source data has not received sufficient attention. To this end, we propose the method of Sparse Common and Distinctive Covariates Logistic Regression (SCD-Cov-logR). The method is a multi-source extension of principal covariates regression (PCovR) [1] that combines with generalized linear modeling framework to allow classification of a categorical outcome. PCovR is a dimension reduction method that extracts components that explain the xvariance in both predictor and outcome variables. In a simulation study, SCD-Cov-logR resulted in outperformance compared to related methods commonly used in behavioural sciences. We also demonstrate the practical usage of the method under an empirical dataset.

Keywords: dimension reduction, logistic regression, multiblock data, principal covariates regression

References

1. De Jong, S., Kiers, H.A.: Principal covariates regression: part i. theory. Chemom. Intell. Lab. Syst. **14(1-3):** 155–164 (1992)

Soogeun Park, Katrijn Van Deun
Tilburg University, Department of Methodology and Statistics, Simon Building, Prof. Cobbenhagenlaan 225, 5037DB Tilburg, e-mail: s.park_1@tilburguniversity.edu

Eva Ceulemans
KU Leuven, Kwantitatieve Psychologie en Individuele Verschillen, Tiensestraat 102 - bus 3713, 3000 Leuven

Accuracy Measures for Binary Classification Based on Quantitative Group Tests

Rui Santos, João Paulo Martins, and Miguel Felgueiras

Classification of a large number of individuals using individual tests can be expensive and time-consuming. Hence, taking a sample from different individuals and mixing it into a homogeneous fluid (the pooled sample) may be a methodology to be taken into account. Based on quantitative group testing, different classification procedures can be performed to save resources, although the probability of misclassification may increase due to the dilution of the discriminant substance in the pooled sample. In this work, the specificity (φ_e) and sensitivity (φ_s) of a classification methodology based on quantitative group tests are used to create a ROC curve which depends on the sensitivity and specificity from individual and group tests (i.e., from the cut-off points applied in the individual and in the group tests), as well as on the group size. These ROC curves are applied to assess the reliability of classification of some methodologies based on quantitative group tests. The results were computed by simulation using populations with 10^6 individuals, different distributions for the discriminant substance (Gaussian, Weibull, Pareto, Lévy, among others) setting different measures for the quality of the individual tests $\varphi_s = \varphi_e \in \{0.99, 0.95, 0.9, 0.8, 0.7\}$, prevalence rates $p \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$ and group sizes $n \in \{2, 5, 10, 20, 50, 100\}$. In some cases, it is possible to almost maintain the accuracy of the individual test and achieve a significant gain in efficiency.

Keywords: classification, dilution effect, group test, Roc curve, simulation

References

1. Hughes-Oliver, J.: Pooling experiment for blood screening and drug discovery, screening methods for experimentation in industry, Drug Discovery and Genetics, Springer, 48–68 (2006)
2. Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, D.: Comparison of group testing algorithms for case identification in the presence of testing error, Biometrics **63**, 1152–63 (2007)
3. Santos, R., Martins, J.P., Felgueiras, M.: An Overview of Quantitative Continuous Compound Tests. In Bourguignon, J.P., Jeltsch, R., Pinto, A., Viana, M. (eds.) Dynamics, Games and Science, CIM Series in Mathematical Sciences **1**, 627–641 (2015)
4. Santos, R., Martins, J., Felgueiras, M., Ferreira, L.: Accuracy Measures for Binary Classification Based on a Quantitative Variable, REVSTAT-STAT J **17**(2), 223–244 (2019)

Rui Santos and Miguel Felgueiras
School of Technology and Management, Polytechnic Institute of Leiria, CEAUL – Center of Statistics and Applications, e-mail: rui.santos@ipleiria.pt,mfelg@ipleiria.pt

João Paulo Martins
School of Health, P. Porto, CEAUL – Center of Statistics and Applications,
e-mail: jom@ess.ipp.pt

Exploiting Pareto Density Estimation for Nonparametric Naïve Bayes Classifiers

Quirin Stier and Michael C. Thrun

In parametric Naïve Bayes classifiers, a variety of class conditional distributions are defined if prior knowledge about the structures in data is given. Otherwise, likelihood estimation is performed via kernel density estimation in non-parametric Naïve Bayes classifiers. However, our previous work showed that Pareto density estimation (PDE) [1] outperforms other density estimation methods available in R and Python, because conventional methods pose several problems when estimating distributions that have clipped data or are uniform, multimodal or skewed [1]. In contrast, PDE is particularly suitable for discovering structures in continuous data and allows for the discovery of mixtures of Gaussians [2]. This work proposes a non-parametric Naïve Bayes classifier called PDEbayes that estimates the likelihood per class via PDE. It is compared with a non-parametric Naïve Bayes classifier available on CRAN called *naivebayes* on a range of artificial datasets of the FCPS ($N = 1000$ samples) [3]. Moreover, a real-world dataset is used which consists of patients with either a positive B-Non-Hodgkin lymphoma (B-NHL) or a negative B-NHL diagnosis for which no prior knowledge about the distributions is available. PDEbayes outperforms the Naïve Bayes classifier on FCPS datasets slightly and on the real-world dataset significantly with a precision of 86% and recall of 85% for PDEbayes, 82% and 76% for naivebayes for $N = 19135$ testdata patients.

Keywords: kernel density estimation, Bayes, classification

References

1. Thrun, M. C., Gehlert, T., Ultsch, A.: Analyzing the Fine Structure of Distributions, PloS one, Vol. 15(10), (2020)
2. Ultsch, A., Thrun, M. C., Hansen-Goos, O., Lötsch, J.: Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss), International Journal of Molecular Sciences, Vol. 16(10), (2015)
3. Thrun, M. C., Ultsch, A.: Clustering Benchmark Datasets Exploiting the Fundamental Clustering Problems, Data in Brief, Vol. 30(C), (2020)

Quirin Stier
IAP-GmbH Intelligent Analytics Projects, Adelheidsdorf, Germany,
e-mail: q.stier@iap-gmbh.de

Michael C. Thrun
Dept. of Mathematics and Computer Science, Philipps-University, Marburg, Germany
e-mail: m.thrun@informatik.uni-marburg.de

Author Index

Özge Sahin, 32
Éric Beaudry, 179
Łukasz Smaga, 96

Adalbert Wilhelm, 115, 153
Adelaide Figueiredo, 136, 140
Adelaide Freitas, 226, 269
Adi Lausen, 120, 227, 266
Adilson Xavier, 86
Aeyeon Lee, 19
Afshin Ashofteh, 252
Agustín Mayo-Iscar, 48, 74, 105, 239, 240
Alba M. Franco-Pereira, 101
Albert Meco, 190
Alejandro Chacón, 160
Aleksandr Koshkarov, 109
Aleksandra Łuczak, 215, 216
Alessandra Menafoglio, 98
Alessandro Bitetto, 126
Alessia Pini, 218
Alessio Farcomeni, 91
Alessio Squassina, 117
Alex Cucco, 242
Alex Partner, 227
Alexandre Penha, 262
Alexei Vernitski, 227
Alfonso Iodice D'Enza, 23, 131
Ali Mertcan Kose, 265
Alice Barth, 268
Alicia Nieto-Reyes, 217
Almond Stöcker, 71
Alya Alzahrani, 93
Amirah S. Alharthi, 92
Amovin-Assagba Martial, 77
Amy LaLonde, 235
An-Chiao Liu, 168
Ana Alexandra Martins, 250
Ana Catarina Nunes, 259
Ana Helena Tavares, 201
Ana Julieta Morais, 226
Ana Margarida Bento, 274, 275
Ana Maria Rodrigues, 259
Ana Moreno, 263
Ana Teles-Machado, 263
Anabela Afonso, 87, 169, 207
Anabela Rocha, 226
Andre C.P.L.F. de Carvalho, 214
Andrea Cappozzo, 42, 105
Andrea Cerioli, 48
Andrea Diana, 142
Andrea Papola, 199
Andreas Artemiou, 93
Andrej Svetlošák, 60
Andrzej Dudek, 150
Andrzej Sokolowski, 94
Angela Montanari, 2, 25
Angeline Plaud, 251
Angelos Markos, 23, 131
Aniek Sies, 118

Anja Ernst, 270
Anna Maria Paganoni, 154, 221
Anna Meloni, 117
Annabella Astorino, 148, 181
Anthony Bagnall, 18
Antonio Balzanella, 51
Antonio D'Ambrosio, 113
Antonio Elías, 102, 221
Antonio Fernández-Anta, 41
Antonio Fuduli, 148, 180, 181
Antonio Irpino, 51, 188, 190
Antonio Pellicani, 57
Antonio Punzo, 44
António Grilo, 262
Anuradha Roy, 81
Arlete Rodrigues, 158
Astrid Dürauer, 264
Aylin Yaman Kocadağlı, 68

Barbara Batóg, 135
Barbara Hammer, 198
Barbara Japelj Pavešić, 50
Barbara Pawelek, 254
Bart Jan van Os, 119
Beata Bal-Domańska, 161
Beate Jahn, 115
Belén Pulido, 101
Bernd Bischl, 30
Berthold Lausen, 120, 203, 224, 225, 227, 266
Bertus Jeronimus, 270
Bettina Grün, 125, 236
Bo Peng, 258
Bogdan Mazoure, 222
Boris Beranger, 189
Boris Mirkin, 246
Brendan McCabe, 127
Brígida Mónica Faria, 114
Byungtae Seo, 232
Bárbara Pereira, 206

Cajo J. F. ter Braak, 39
Carla Henriques, 208
Carlo Cavicchia, 23
Carlos Matrán Bea, 238
Carlos Soares, 158, 214
Carolina Cardoso, 208
Casper Albers, 270
Catarina Marques, 211
Caterina Gregorio, 257
Cecilia Salvatore, 149
Charles Bouveyron, 11
Charles C. Taylor, 92
Chiara Masci, 154
Chin Pang Ho, 184
Chris Saker, 227
Christian Hennig, 237
Christian Riccio, 199
Christine Keribin, 183
Christophe Biernacki, 183
Chun-Yang Peng, 156
Cinzia Di Nuzzo, 85
Cinzia Viroli, 91
Clara Yokochi, 262
Claudia Czado, 32
Claudia Kirch, 217
Claudia Nunes Philippart, 76
Claudia Pisanu, 117
Claudio Agostinelli, 47
Claudio Conversano, 122
Cláudia Silvestre, 45
Cristina Anton, 73
Cristina Lopes, 259
Cristina Oliveira, 259
Cristina Tortora, 75, 155

Da Hee Oh, 172
Daniel Peña, 16
Daniel Santos, 169, 207
Daniela Silva, 263
David Masís, 86
David P. Hofmeyr, 261
David Rodríguez Vítores, 238
Deborah Rohm Young, 235
Delia Francesca Chillura Martino, 72
Dianne Cook, 10
Diogo Alves, 129
Diogo Pinheiro, 152
Dmitry Frotov, 246
Dolores Romero Morales, 147

Dominique Desbois, 151
Dongha Kim, 84
Douglas C. Montgomery, 156
Duarte Rodrigues, 114
Düzgün Yıldırım, 63

Edoardo Redivo, 91
Eduardo Andre Costa, 137
Edward A. Baron, 274
Edwin Diday, 4, 50
Elżbieta Sobczak, 162
Eleonora Arnone, 176
Elia Cunial, 176
Elif Göksu Öztürk, 259
Elisabete Teixeira, 275
Elise Dusseldorp, 119
Elvira Romano, 142
Emilio Carrizosa, 147
Engelbert Mephu Nguifo, 205, 251
Eri Hoshino, 82
Eric J. Beh, 7, 173
Erick Chatalov, 206
Erika Nakanishi, 82
Erin I. McDonnell, 255
Esteban Segura, 86
Ester Zumpano, 180
Etienne Low-Decarie, 266
Eugenio Vocaturo, 180
Eun-Kyung Lee, 21
Eva Ceulemans, 276
Ewa Genge, 67
Eyke Hüllermeier, 197, 198

Fabian Fumagalli, 198
Fabian Scheipl, 43
Fabien Llobell, 46
Fabrizio Maturo, 219
Fatemeh Asgari, 178
Fatma Sevinç Kurnaz, 78
Federico D'Onofrio, 260
Fei Liu, 52
Felix Gnettner, 217
Feng Ji, 270
Fernanda Figueiredo, 136, 140
Fernanda Sousa, 166
Fernando Silva, 128
Filipe J. Marques, 262
Filipe Magalhães, 166
Filippo Antonazzo, 159, 183
Flora Ferreira, 166
Florence Forbes, 182
François Bavaud, 95
Francesca Chiaromonte, 141
Francesca Di Salvo, 72
Francesca Greselin, 42, 105
Francesca Ieva, 154, 257
Francesco Bartolucci, 67, 271
Francesco Denti, 42
Francesco Palumbo, 155, 229
Francisca G. Vieira, 262
Friedrich Leisch, 264
Fulvia Pennoni, 271
Fumio Ishioka, 191

Gabriel Martos Venturini, 60, 89
Gabriele Perrone, 104
Gabriele Soffritti, 104
Gabriella Chirco, 72
Gannaz Irène, 77
Genevera I. Allen, 12
Georgia Panagiotidou, 69
Gerard B.M. Heuvelink, 261
Germán Aneiros, 177
Gero Szepannek, 153
Gersende Fort, 182
Giancarlo Ragozini, 56
Gianluca Morelli, 74, 240
Gianpaolo Zammarchi, 122
Gianvito Pio, 57
Giorgia Zaccaria, 267, 269
Giorgio Grani, 260
Giovanni Saraceno, 47
Giulia Barbati, 257
Giuliano Galimberti, 25
Giuseppe Giordano, 56
Glòria Mateu-Figueras, 99, 134, 170
Gonçalo Jacinto, 87, 169, 207
Guojun Gan, 220
Guyslain Naves, 194

Hae-Hwan Lee, 55

Hana Řezanková, 33
Hans-Peter Piepho, 165
Helena Bacelar-Nicolau, 59, 223
Henk A.L. Kiers, 123
Henrik Nordmark, 203
Henrique Siqueira, 157
Herbert K. H. Lee, 65
Heungsun Hwang, 230
Hien Duy Nguyen, 182
Hiroshi Yadohisa, 66, 192
Ho Kim, 20
Hong Kyu Lee, 167
Hui Yang, 224, 225
Hulin Wu, 220
Hung Tong, 75
Hyunjoong Kim, 53
Hyunsuk Kim, 233
Hélder Sousa, 275

Iain Smith, 73
Igor Kravchenko, 152
Ilsu Choi, 231
Immanuel M. Bomze, 61, 258
Ines Wilms, 133
Inyoung Kim, 230
Inês Oliveira e Silva, 158
Isabel Pereira, 127
Isabel Ribeiro, 166
Isabel Silva, 127
Iva Karafiátová, 273
Iven Van Mechelen, 118
Ivone Figueiredo, 206

Jaël Champagne Gareau, 179
Jacek Batóg, 135
Jacques Julien, 77
Jaesung Hwang, 84
Jan Graffelman, 40, 99
Jan Kalina, 187
Jan Michael Spoor, 247
Jangsun Baek, 83
Jaqueline Meulman, 119
Jasminka Dobša, 123
Jasone Ramírez-Ayerbe, 147
Javier Arroyo, 190
Javier Palarea-Albaladejo, 133, 134
Javier Trejos, 86, 160
Jayoeng Paek, 231
Jean Diatta, 196
Jean-Marc Ferrandi, 46
Jeff Goldsmith, 255
Jeffrey Durieux, 34
Jenni Niku, 174
Jennifer A. Schrack, 255
Jens Weber, 247
Jiří Dvořák, 273
Jinwon Heo, 83
Jivka Ovtcharova, 247
João Alves, 61
João Gama, 6, 13
Joachim Behnke, 115
Joachim Engel, 115
Joerg Blasius, 139
Johannes Fürnkranz, 29, 197
Johané Nienkemper-Swanepoel, 22
Jongho Im, 55
Joni Virta, 62
Jorge Arce Garro, 79
Jorge Mateu, 142
José Soares, 166
José Campos e Matos, 274
José Soeiro Ferreira, 259
Jose Luis Vicente-Villardon, 38, 175
Josefa E. Großschedl, 61
Josep A. Martín-Fernández, 133, 134, 170
Joseph E. Yukich, 102
Joseph Ogutu, 165
Joshua Tobin, 184
José A. Vilar, 17, 248
José Campos e Matos, 275
José Matos, 129
José Pedro Lopes, 262
José Saias, 169, 207
Jozef Pociecha, 254
João Lagarto, 250
João Paulo Martins, 277
João Paulo Rodrigues, 262
Ju-Young Park, 232
Juan C. Vera, 110

Juan Claramunt Gonzales, 119
Juan Pablo Equihua, 203
Julia Wrobel, 255
Julian Rossbroich, 34
Jun Li, 19
Jun Tsuchida, 66, 192
Junji Nakano, 24

Karel Hron, 133, 170
Karolina Pokorska, 162
Kateřina Pawlasová, 273
Katerina M. Marcoulides, 272
Katrijn Van Deun, 110–112, 168, 276
Kenneth Gavin, 274
Klaas Sijtsma, 110
Klaus Nordhausen, 62
Koji Kurihara, 191
Kotomi Sakai, 82
Kristofer Bouchard, 65
Kuniyoshi Hayashi, 82
Kyeongah Nah, 231

Lan Liang, 99
Lara Fontanella, 242
Laura Anderlucci, 25
Laura M. Sangalli, 98, 176, 221
Laura Palagi, 260
Laura Trinchera, 272
Laura Vicente-Gonzalez, 175
Lazhar Labiod, 106, 107, 210, 241
Leonor Rego, 87, 169, 207
Liming Liang, 19
Lina Oliveira, 152
Lisa Steyer, 71
Lisete Sousa, 202, 206
Louis Tran, 75
Luca Bagnato, 44
Luca Greco, 47
Lucio Barabesi, 48
Lucio Palazzo, 229, 249
Luigi Ippoliti, 100
Luis A. García-Escudero, 48, 74, 105, 239, 240
Luis Paulo Reis, 114
Luísa Novais, 164
Lynne Billard, 52

M. Graça Batista, 223
Maciej Łuczak, 88
Magdalena Talaga, 244
Maged Ali, 203
Malgorzata Markowska, 94
Man-Suk Oh, 167
Manuel Rui Alves, 132
Manuela Schmidt, 124
Marc Comas-Cufí, 134
Marc Ditzhaus, 96
Marcela Zembura, 244
Marcella Niglio, 64
Marcelo Silva, 87, 169, 207
Marcin Pelka, 150, 162, 188
Marco Alfò, 26
Marco Riani, 74, 240
Margarida G. M. S. Cardoso, 45, 250
Maria de Fátima Salgueiro, 211
Maria do Rosário Oliveira, 76, 152
Maria Eduarda Silva, 127, 128, 137
Maria Margarida Lima, 259
Maria Prosperina Vitale, 56
Marialuisa Restaino, 64
Marieke Timmerman, 270
Marilia Antunes, 202
Marjolein Fokkema, 121
Markus Pauly, 115
Markus Zwick, 115
Marta Belchior Lopes, 200, 262
Marta Monaci, 260
Marta Nai Ruscone, 113
Maryam Al Alawi, 103
Marzia A. Cremona, 141
Masayuki Obatake, 82
Matteo Avolio, 148, 180, 181
Matteo Farnè, 25, 90, 204
Matteo Pegoraro, 70
Matthias Medl, 264
Matthieu Resche-Rigon, 195
Maurizio Romano, 122
Maurizio Vichi, 269
Mauro Iannuzzi, 204
Maximilian Muschalik, 198

Mayetri Gupta, 103
Mayumi Tanahashi, 193
Małgorzata Just, 216
Metodi Metodiev, 120
Michael C. Thrun, 35, 278
Michael Dinzinger, 205
Michael Fop, 26, 209
Michael Franklin Mbouopda, 205
Michael P. B. Gallaugher, 145, 146
Michael Rapp, 197
Michal Swachta, 188
Michel van de Velden, 23, 131
Michelangelo Ceci, 57
Michele La Rocca, 64
Michele Staiano, 199
Michelle Carey, 220
Miguel de Carvalho, 60, 89
Miguel Felgueiras, 277
Mikhaël Carmona, 194
Mimi Zhang, 184
Min Soo Kim, 54
Minh Quang Tran, 275
Mitsuyoshi Suzuki, 82
Mohamed Achraf Bouaoune, 222
Mohamed Hanafi, 46
Mohamed Nadif, 106, 107, 210, 241
Moises Santo, 214
Moritz Herrmann, 43
Mário Figueiredo, 45

Nadia Tahiri, 109, 222
Nail Chabane, 222
Ndèye Niang, 195
Neslihan Gökmen İnan, 63, 68
Nicola Pronello, 100, 242
Niels Lundtorp Olsen, 218
Nihan Acar-Denizli, 256
Niël le Roux, 22
Nobuo Shimizu, 24
Nosheen Faiz, 120
Nádia Bachir, 262

Olaf Wolkenhauer, 115
Oldemar Rodríguez Rojas, 79, 80
Olivier Cappé, 182
Oluwasegun T. Ojo, 41
Onay Urfalioglu, 157
Osvaldo Silva, 59, 223
Ozan Kocadağlı, 63, 68, 265

Pablo Montero-Manso, 17
Paola Cerchiello, 126
Pascal Préa, 194
Patrice Bertrand, 196
Patrick Groenen, 130
Patrik Janáček, 187
Patrícia Gois, 169, 207
Paul D. McNicholas, 145, 146
Paul Hofmarcher, 125
Paula Brito, 201
Paula C.R. Vicente, 211
Paulo Infante, 87, 169, 207
Paulo Quaresma, 169, 207
Paulo Rebelo Manuel, 87, 169, 207
Paweł Lula, 244
Paweł Piasecki, 88
Pedro Bastardo, 158
Pedro Campos, 243, 252
Pedro Delicado, 256
Pedro Espadinha-Cruz, 262
Pedro Filipe Rocha, 259
Pedro Nogueira, 87, 169, 207
Pedro Pacheco, 166
Pedro Pinto, 208
Pedro Ribeiro, 128
Pedro Sá Couto, 226
Pepus Daunis-i-Estadella, 170
Peter A. Tait, 145
Peter Filzmoser, 49, 78, 133
Petra Laketa, 186
Philippe Vieu, 177
Piercesare Secchi, 70, 98
Pierpaolo D'Urso, 248
Prosper Ablordeppey, 269

Quirin Stier, 35, 278

Rūta Petraitytė, 224
Rabea Aschenbruck, 153
Rafal Kulakowski, 266
Raffaella Calabrese, 60

Rafik Abdesselam, 245
Rafika Boutalbi, 106, 210
Rainer Dyckerhoff, 185
Ralf Münnich, 115
Raquel Menezes, 263
Rasool Taban, 76
Raúl Jiménez, 102, 221
Reda Amir Sofiane Tighilt, 222
Regina Bispo, 262
Riccardo Giubilei, 108
Riccardo Ievoli, 249
Riccardo Scimone, 98
Rita Coimbra, 171
Roberta Siciliano, 199
Roberto Ascari, 97
Rong Pan, 156
Rosa E. Lillo, 41, 101
Rosalina Pisco Costa, 169, 207
Rosangela Ballini, 253
Rosanna Verde, 51, 219
Rosaria Ignaccolo, 100
Rosaria Lombardo, 7, 173
Rosember Guerra-Urzola, 110
Rosy Oh, 167
Ruey S. Tsay, 16
Rui Calçada, 274
Rui Meng, 65
Rui Rodrigues, 165
Rui Santos, 277
Ruta Petraityte, 225
Ryan P. Browne, 146

Salvatore D. Tomarchio, 144
Salvatore Ingrassia, 85, 159
Samil Uysal, 119
Sander Scholtus, 168
Sandra Silva, 129
Sanetoshi Yamada, 193
Sangwook Kang, 232
Sanjeena Dang (Subedi), 234
Sara Cabral, 223
Sara Fontanella, 100, 242
Sara Taskinen, 174
Sarah Friedrich, 115
Scott Sisson, 189
Sebastian Ratzenböck, 61
Seong Tak Woo, 172
Seunghwan Park, 55
Seungyeoun Lee, 230, 233
Shoji Kajinishi, 191
Shuang Wu, 220
Silvia D'Angelo, 26
Silvia Novo, 177
Silvia Pandolfi, 271
Simona Korenjak-Černe, 50
Simone Vantini, 218
Sofia Magopoulou, 138
Sonia Migliorati, 97
Sonja Greven, 71
Soogeun Park, 276
Sophie Dominique, 46
Sourav Adhikari, 125
Stanislav Nagy, 185, 186
Stefan Meingast, 61
Stefano A. Gattone, 143
Stella Hadjiantoni, 224, 225
Stephan van der Westhuizen, 261
Su Hoon Choi, 54
Sugnet Lubbe, 22
Suhyun Hwangbo, 230
Sungyoung Lee, 230
Surajit Ray, 103, 163
Susana Faria, 164, 213
Susana Garrido, 263
Susana Nascimento, 246
Susana Vinga, 200
Sławomir Kalinowski, 215

Tadashi Imanishi, 193
Taerim Lee, 20, 116
Taesung Park, 19, 230, 233
Takehiro Shoji, 66
Tanzy Love, 235
Thais Pacheco Menezes, 209
Theodore Chadjipadelis, 69, 138
Theresa Scharl, 236, 264
Thomas Brendan Murphy, 209
Thomas Whitaker, 189
Tim Friede, 115
Tobia Boschi, 141

Tom F. Wilderjans, 34
Tomasz Górecki, 88
Tomáš Kliegr, 28
Ton de Waal, 168
Tongtong Wu , 235
Tonio Di Battista, 143
Torsten Möller, 61
Toshiki Sakai, 192
Tra Le, 112
Trevor Fenner, 246
Tüzün Tolga İnan, 68

Una Radojicic, 62
Ursula Garczarek, 115
Urszula Cieraszewska, 244
Utkarsh J. Dang, 146
Uwe Sieber, 115

Vadim Zipunnikov, 255
Valentin Todorov, 49
Valeria Vitelli, 178
Vanda Lourenço, 165
Vanessa Freitas Silva, 128
Vera Afreixo, 201
Veronica Piccialli, 149
Veronne Yepmo, 251
Victor Chepoi, 194
Viktorie Nesrstová, 133, 170
Vincent Audigier, 195
Vincenzo Giuseppe Genova, 56
Vitor Nogueira, 169, 207
Vladimir Batagelj, 58
Vladimir Makarenkov, 179, 222
Volodymyr Melnykov, 27, 212
Véronique Cariou, 46

Wangshu Tu, 234
Wenhui Zhang, 163
Wessel H. Moolman, 228
Wonil Chung, 19

Xuwen Zhu, 212

Yana Melnykov, 212
Yang Wang, 27
Ye-eun Kim, 53
Yongdai Kim, 84
Yongkuk Kim, 231
Yoshikazu Yamamoto, 24
Yoshiro Yamamoto, 193
You-Jin Park, 156
Youngkwang Cho, 19
Youngmi Kim Pak, 167
Yunfei Long, 224, 225

Zardad Khan, 120
Zdeněk Šulc, 33
Zhirayr Hayrapetyan, 246
Zoë-Mae Adams, 22

Ángel López-Oriona, 17, 248
Áurea Sousa, 59, 223

İsmail Meşe, 63

IFCS 2022

17th Conference of the International Federation of Classification Societies

Classification and Data Science in the Digital Age

Book of Abstracts