Book of Abstracts

18th Conference of the International Federation of Classification

Data Science, Classification and Artificial Intelligence for Modeling Decision Making

8th Latin American Conference on Statistical Computing

Recent Advances in Statistical Computing and Data Science

San José, Costa Rica

July 15–19, 2024

Welcome

We are very pleased to welcome you to IFCS 2024, "Data Science, Classification and Artificial Intelligence for Modeling Decision Making". This is the 18th Conference of the International Federation of Classification Societies (IFCS), which will take place from July 15 to 19, 2024 in San José, Costa Rica.

The local organization has been managed by the Center for Research in Pure and Applied Mathematics (CIMPA) of the University of Costa Rica, with support from the Office of the Rector, Office of Social Action Vice-Rector, Office of Research Vice-Rector, School of Mathematics, FundacińUCR, Office of International Affairs, and Office of General Services.

We extend our gratitude to our sponsors, BAC Credomatic, Integro Grupo, and Instituto Nacional de Seguros.

The International Federation of Classification Societies (IFCS) groups the national or regional classification societies all around the world. The IFCS organizes an international conference every two years since 1987.

The IFCS was founded in 1985, it is composed of 18 national, regional or linguistically-based data science and classification societies al around the world. These societies are:

- Associação Portuguesa de Classificação e Análise de Dados (CLAD)
- British Data Science Society (BDSS)
- The Classification Society (CS)
- Gesellschaft für Klassifikation (GfKl)
- Helenic Society for Data Analysis (GSDA)
- Hungarian Statistical Association (HSA-CMSG)
- Irish Pattern Recognition and Classification Society (IPRCS)
- Japanese Classification Society (JCS)
- Korean Classification Society (KCS)
- Moroccan Classification Society (MCSo)
- Multivariate Data Analysis Group of the South African Statistical Association (SASA-MDAG)
- Multivariate Statistics and Classification Group of the Spanish Society of Statistics and Operations Research (SEIO-AMyC)
- Sekcja Klasyfikacji i Analizy Danych PTS (SKAD)
- Sociedad Centroamericana y del Caribe de Clasificación y Análisis de Datos (SoCCCAD)
- Classification and Data Analysis Group of the Italian Statistical Society (ClaDAG-SIS)
- Société Francophone de Classification (SFC)
- Statistično društvo Slovenije (SdS)
- Vereniging voor Ordinatie en Classificatie (VOC)

Periodically, the IFCS organizes conferences where the most recent developments of classification, data science, data analysis, data mining, machine learning and applications using these tools are presented and published. There have been seventeen conferences

1. Aachen, Germany (1987),
2. Charlottesville, USA (1989),
3. Edinburgh, Scotland (1991),
4. Paris, France (1993),
5. Kobe, Japan (1996),
6. Rome, Italy (1998),
7. Namur, Belgium (2000),
8. Cracow, Poland (2002),
9. Chicago, Illinois (2004),
10. Ljubljana, Slovenia (2006),
11. Dresden, Germany (2009),
12. Frankfurt, Germany (2011),
13. Tilburg, The Netherlands (2013),
14. Bologna, Italy (2015),
15. Tokyo, Japan (2017),
16. Thessaloniki, Greece (2019),
17. Porto, Portugal (2022).

Scientific Program Committee

- Javier Trejos (University of Costa Rica), co-chair
- Rebecca Nugent (IFCS President), co-chair
- Angela Montanari (IFCS Past-president), co-chair
- Adalbert Wilhelm (Constructor U, Germany, GfKl),
- Atsuho Nakayama (Doshisha U, Japan, JCS),
- Aurea Grané (SEIO-AMyC),
- Balázs Horváth (U Eötvös Loránd, Hungary, MST),
- Berthold Lausen (Essex U, United Kingdom, BDSS),
- Carlos Cuevas-Covarrubias (U Anáhuac, Mexico, SOCCAD),

- Hyunjoong Kim (Yonsei U, Korea),
- Jean Diatta (U La Réunion, France, SFC),
- Johané Nienkemper-Swanepoel (Stellenbosch University, South Africa, SASA-MDAG),
- Krzysztof Jajuga (Wroclaw U of Economics and Business, Poland, SKAD),
- Mark de Rooij (Leiden U, The Netherlands, VOC),
- Maurizio Vichi (U La Sapienza Roma, CLADAG),
- Michael Gallaugher (Baylor U, USA, TCS),
- Paula Brito (U Porto, Portugal, CLAD),
- Simona Korenjak-Černe (Ljubljana U, Slovenia, STAT), and
- Sonya Coleman (Ulster U, North Ireland, IPRCS),
- Theodore Chadjipadelis (GSDA).

Local Organizing Committee

- Adriana Sánchez (Mathematics, UCR), co-chair,
- Alex Murillo (Atlantic Campus, UCR),
- Allan Berrocal (Computer Science, UCR),
- Ana María Durán (Physics, UCR),
- Edgar Casasola (Computer Science, UCR),
- Fabio Sanchez (Mathematics, UCR),
- Javier Trejos (Mathematics, UCR), chair,
- Jorge Arce (Computer Science, UNA),
- Juan Gabriel Calvo (Mathematics, UCR),
- Juan José Leitón (National Institute of Electricity),
- Luis Amaya (Guanacaste Campus, UCR),
- Luis Barboza (CIMPA, UCR),
- Mario Villalobos (Mathematics, UCR), co-chair,
- Marvin Coto (Electric Engineering, UCR),
- Minor Bonilla (Grupo Montecristo), and
- Shu-Wei Chou (Statistics, UCR).
- María Luisa González (CIMPA), manager,

- Felipe Escalante (CIMPA), webmaster,
- José Luis Amador (Mathematics), web assistant,
- Alejandro Mairena (CIMPA), OCS assistant,
- María Paula Brenes (Mathematics), edition assistant,
- Melissa Arguedas (Mathematics), edition assistant,
- Jorge Mario Valverde Ramírez (CIMPA), edition assistant.

Sponsors

BAC Credomatic
Intego Group LLC
National Insurance Institute (INS)
University of Costa Rica (UCR)

Partners

Faculty of Engineering, UCR
Faculty of Science, UCR
FundaciónUCR (UCR Foundation)
International Affairs and External Cooperation Office, UCR
International Association for Statistical Computing (IASC)
International Federation of Classification Societies (IFCS)
Latin American Regional Section (LARS) of IASC
Office for General Services, UCR
Rector Office and Social Action Vice-Rector Office, UCR
School of Mathematics, UCR
Springer

Organization

Central American and Caribbean Society for Classification and Data Analysis (SoCCAD)
Research Center for Pure and Applied Mathematics (CIMPA), UCR

Program

Monday, 15

8:00 – ∞: Registration: `UCR` .

08:30 Session: Tutorial 1 (Tut01): `Room 1`.

08:30 – 10:00 ALFARO, M.: Reproducible data analysis (pag. 38).

08:30 Session: Tutorial 2 (Tut02): `Room 2`.

08:30 – 10:00 MATABUENA, M.: Statistical Science Meets Digital Health (pag. 101).

10:00 – 10:30 : Coffee break.

10:30 Session: Tutorial 1 (Tut01)(Cont...): `Room 1`.

10:30 – 12:00 ALFARO, M.: Reproducible data analysis (pag. 38).

10:30 Session: Tutorial 2 (Tut02)(Cont...): `Room 2`.

10:30 – 12:00 MATABUENA, M.: Statistical Science Meets Digital Health (pag. 101).

12:00 – 13:30 Tiempo para almuerzo / Time for lunch.

14:00 – 14:30 : Opening ceremony `Auditorium`.

14:30 Session: Opening Plenary Talk (Conf1): `Auditorium`.

14:30 – 15:30 GROENEN, P.: MM-Algorithms in Data-Science (pag. 76).

15:30 – 16:00 : Coffee break.

16:00 Session: Modeling Multivariate Data (A. Roy) (MMD1-1): `Room 1`.

16:00 – 16:20 MARTINEZ, A.: Multiblock Methods for Learning Structural Equation Models: An Overview (pag. 99).

16:20 – 16:40 FRANCZAK, B.: Classifying multivariate observations in data sets with asymmetric features and outlying observations (pag. 72).

16:40 – 17:00 SHULTS, J.: Accounting for the Shutdown Due to the COVID Pandemic in an Analysis of Multivariate Data from a School and Medical Practice-Based Intervention: the West Philadelphia Asthma Care Implementation Study (pag. 136).

17:00 – 17:20 SAJOBI, T.: A comparison of multivariate mixed models and generalized estimation equations models for discrimination in multivariate longitudinal data (pag. 130).

16:00 Session: Symbolic Data Analysis 3 (SymDA3-1): `Room 2`.

16:00 – 16:20 BRITO, P. & DIAS, S. & NIANG, N.: Quality Measures for Clusterwise Regression (pag. 53).

16:20 – 16:40 CHAPARALA, P.: Symbolic Data Analysis Framework for Recommendation Systems: SDA-RecSys (pag. 61).

16:40 – 17:00 TAHIRI, N. & FARNIA, M.: A new metric to classify B cell lineage tree (pag. 142).

17:00 – 17:20 CROCETTA, C. & IRPINO, A: New tools for visualizing distributional datasets (pag. 66).

16:00 Session: Visualization (J. Nienkemper & S. Lubbe) (MD_V1-1): `Room 3`.

16:00 – 16:20 DE-ROOIJ, M.: Reduced Rank Regression with Mixed Predictors and Mixed Responses (pag. 68).

16:20 – 16:40 SCHOONEES, P.: Model Selection for Linear Regression Under Data Aggregation (pag. 135).

16:40 – 17:00 CAVICCHIA, C. & IODICE D'ENZA, A. & VAN, M.: Nearest neighbors for mixed type data: an inter-dependency based approach (pag. 54).

17:00 – 17:20 NAKAYAMA, A.: Integration of Deep Learning and Marketing Research for Brand Confusion Prediction and Visual ad Analysis (pag. 111).

17:25 Session: Dimension Reduction (DimREd1-1): `Room 1`.

17:25 – 17:45 GUERRA, R.: Optimal penalized sparse PCA (pag. 77).

17:45 – 18:05 LE, T.: Multiblock Regularized Least-squares Latent Variable Method (pag. 86).

17:25 Session: Optimization in Classification and Clustering (OpCC1-1): `Room 2`.

17:25– 17:45 CHAMPAGNE GAREAU, J. & BEAUDRY, E. & MAKARENKOV, V.: Towards Topologically Diverse Probabilistic Planning Benchmarks: Synthetic Domain Generation for Markov Decision Processes (pag. 60).

17:45– 18:05 SOW, K. & GHAZALLI, N.: Machine Learning-Based Classification and Prediction to Assess Corrosion Degradation in Mining Pipelines (pag. 138). (ZOOM session)

17:25 Session: Big Data and High-Dimensional (BD_HD1-1): `Room 3`.

17:25– 17:45 SOLÍS, M. & HERNÁNDEZ, A.: UMAP projections and the survival of empty space: A geometric approach to high-dimensional data (pag. 137).

17:45– 18:05 LOPES, M.: Understanding omics links behind glioma heterogeneity: a network and clustering approach (pag. 88).

19:00 – ∞ : Conference Icebreaker Cocktail

Tuesday, 16

8:00 – ∞: Registration.

08:30 Session: Clustering, Classification and Discrimination 1 (CC_D1-1): `Room 1`.

08:30 – 08:50 COSTA, E. & PAPATSOUMA, I. & MARKOS, A.: A Deterministic Information Bottleneck Method for Clustering Mixed-Type Data (pag. 65).

08:50 – 09:10 ANDERLUCCI, L. & MONTANARI, A.: Randomly perturbed random forests (pag. 40).

09:10 – 09:30 NIANG, N.: Weighted Consensus Clustering for Unbiased Feature Importance in Random Forests (pag. 112).

08:30 Session: Symbolic Data Analysis 1 (SymDA1-1): `Room 2`.

08:30 – 08:50 BILLARD, L. & PAN, W.: Distributional-based Partitioning with Copulas (pag. 49).

08:50 – 09:10 TENORIO, F.: A fuzzy clustering algorithm with entropy regularization for interval-valued data (pag. 144).

09:10 – 09:30 ROY, A. & MONTES, F.: Hypothesis Testing of Mean Interval for p-dimensional Interval-valued Data (pag. 129).

08:30 Session: LACSC: Data Science and Computational Statistics: Theory and applications (LACSC-IS-1): `Room 3`.

08:30 – 08:50 MARTÍNEZ, A. & NIANG, N & LEMUS, P.: Alternating Least Squares Algorithm: Speedup versus Accuracy (pag. 100).

08:50 – 09:10 ACOSTA, J. & ACOSTA, J. & ELLISON, A. & DE CASTRO, M.: Comparing two spatial variables with the probability of agreement (pag. 36).

09:10 – 09:30 NICOLIS, O. & VARINI, E. & ROTOINDI, R. & CAMPUSANO, E. & PERALTA, B. & RUGGERI, F.: Comparing precursors for earthquake prediction in Chile (pag. 113).

9:30 – 10:00 : Coffee break.

10:00 Session: Clustering, Classification and Discrimination 2 (CC-D2-1): `Room 1`.

10:00 – 10:20 FALIH, I. & GONDECH, A. & GOBLET, X. & TRAN, H.: Multimodal Emotion Recognition: A comparative study (pag. 69).

10:20 – 10:40 POOLEY, J. & LAUSEN, B. & MAHMOUD, O.: P-Value Adjusted Selected Tree Ensemble (pag. 121).

10:40 – 11:00 ORTEGA, L.: A toolbox for clustering ordinal data in the presence of missing values (pag. 115).

10:00 Session: Symbolic Data Analysis 2 (SymDA2-1): `Room 2`.

10:00 – 10:20 MATTERA, R. & FRANSES, P.: Spatio-temporal hierarchical clustering of interval time series with application to suicide rates in Europe (pag. 103).

10:20 – 10:40 ARCE, J.: Principal Components Analysis of Histogram-valued Data: Set Theory Approach (pag. 42).

10:40 – 11:00 VERDE, R. & BORRATA, G. & BALZANELLA, A. & DE CARVALHO, F.: A Robust approach of the Clusterwise Regression method for distributional data (pag. 145).

10:00 Session: LACSC session 1: Data Science (LACSC1-1): `Room 3`.

10:00 – 10:20 ZEVALLOS, M & RICARDI, R.: Forecasting Realized Volatility: Does Anything Beat Linear Models? (pag. 148).

10:20 – 10:40 CEVALLOS, H.: Robustness under Missing Data: A Comparison with Special Attention to Inference (pag. 56).

10:40 – 11:00 FALLAS, J.: Comparison of cooling schedules in the simulated annealing algorithm applied to the clustering problem (pag. 70).

11:10 Session: Plenary Talk 2 (Conf2): `Auditorium`.

11:10 – 12:10 CHEN, R.: Bayesian Selection Approach for Categorical Responses via Multinomial Probit Models (pag. 62).

12:10 – 13:30 Time for lunch.

13:30 Session: Plenary Talk 3 (Conf3): `Auditorium`.

13:30 – 14:30 MATABUENA, M.: Statistical Science Meets Digital Health (pag. 102).

14:30 Session: Data Science (DS2-1): `Room 1`.

14:30 – 14:50 RODRIGUEZ, O.: Riemannian Statistics for Any Type of Data (pag. 127).

14:50 – 15:10 BATAGELJ, V. & FERLIGOJ, A.: Network analysis approach to the analysis of event sequences (pag. 48).

15:10 – 15:30 SUBEDI, S. & DANG, U.: Gaussian mixture models for changepoint detection (pag. 140).

14:30 Session: Advances in supervised classification (AdvSCl1-1): `Room 2`.

14:30 – 14:50 SUWANWONG, A. & HARRISON, A. & MAHMOUD, O.: A Gene Selection Method for Classification with Three Classes Using Proportional Overlapping Scores (pag. 141).

14:50 – 15:10 BANKS, D.: Statistics in the Knowledge Economy (pag. 47).

15:10 – 15:30 BOMZE, I. & D'ONOFRIO, F. & PENG, B. & PALAGI, L.: Scalable conic optimization for feature selection in linear SVMs with cardinality control (pag. 51).

14:30 Session: Applications (Apl1-1): `Room 3`.

14:30 – 14:50 VILLALOBOS, K.: TabText: A Flexible and Contextual Approach to Tabular Data Representation (pag. 147).

14:50 – 15:10 MEDL, M.: Optimization Strategies for Bioprocess Parameterization: A Comparative Evaluation (pag. 105).

15:10 – 15:30 CERVANTES ARTAVIA, J. & MONGE CORDONERO, M. & SABATER GUZMÁN, D.: Predicting Air Pollution in Beijing, China Using Chemical, and Climate Variables (pag. 55).

15:30 – 16:00 : Coffee break.

16:00 Session: Plenary Talk 4 (Conf4): `Auditorium`.

16:00 – 17:00 COBO, B.: Integration of data from probability and non-probability surveys (pag. 63).

17:00 Session: Data Mining (DatMin1-1): `Room 1`.

17:00 – 17:20 MEPHU, E. & MAUREEN, N. & JERRY, L: Bridge the gap between Gradual Patterns and Statistical Correlations (pag. 107).

17:20 – 17:40 FRANCE, S.: Visualization and Clustering with Projective Techniques (pag. 71).

17:40 – 18:00 ADHIKARI, S.: Combining Topic Modeling and Word Embedding to Predict Match Outcomes in Association Football (pag. 37).

17:00 Session: Time Series Analysis and Pattern Recognition (TimeSAA-1): `Room 2`.

17:00 – 17:20 SALNIKOV, D. & NASON, G. & CORTINA-BORJA, M.: Modelling clusters in network time series with an application to presidential elections in the USA (pag. 131).

17:20 – 17:40 KHISMATULLINA, M.: Multivariate clustering of nonparametric time trends (pag. 81).

17:40 – 18:00 ASSE, J.: Pattern Recognition for Mexican Household Power Demand Time Series (pag. 44).

17:00 Session: Classification Methods for Large Datasets (A. Grané) (CM-LDA1-1): `Room 3`.

17:00 – 17:20 BOJ, E. & GRANÉ, A. & MAYO-ÍSCAR, A.: Robust distance-based generalized linear models: A new tool for classification (pag. 50).

17:20 – 17:40 GONZALEZ, P.: High-dimensional survival analysis: exploring Cox regression with lasso and adaptive lasso penalties (pag. 74).

17:40 – 18:00 MORALA, P.: Using polynomials to explain classification outputs from neural networks (pag. 109).

18:30 : IFCS Council Meeting – Hotel Aurola

Wednesday 17

08:30 Session: Clustering, Classification and Discrimination 5 (A. Grané) (CC_DA1-1): `Room 1`.

08:30 – 08:50 GRANÉ, A. & SALINI, S. & INFANTE, G.: A new distance for categorical data with moderate association (pag. 75).

08:50 – 09:10 MARTIN, J.: Unsupervised methods for the creation of orthonormal bases in compositional data: R-mode clustering (pag. 95).

09:10 – 09:30 STIER, Q.: An efficient multicore CPU implementation of the DatabionicSwarm (pag. 139).

08:30 Session: Data Science 1 (DSc1-1): `Room 2`.

08:30 – 08:50 PANAGIOTIDOU, G. & CHADJIPADELIS, T.: Mapping Electoral Behavior and Political Competition: A Comparative Analytical Framework for Voter Typologies and Political Discourses (pag. 117).

08:50 – 09:10 BOURANTA, V. & PANAGIOTIDOU, G. & CHADJIPADELIS, T.: Candidates, Parties, Issues and the Political Marketing Strategies: A Comparative Analysis on political competition in Greece (pag. 52).

09:10 – 09:30 CORRALES-BARQUERO, R.: Gender Bias Mitigation in a Credit Scoring Model (pag. 64).

08:30 Session: LACSC session 3: Applications (LACSC3-1): `Room 3`.

08:30 – 08:50 GUTIERREZ, E. & CHOU-CHEN, S. & SOMARRIBAS-BLANCO, M.: Analysis of Intoxication Cases Reported in Costa Rica from 2020 to 2022: Before and during the COVID-19 Period. (pag. 79).

08:50 – 09:10 REDIVO, E. & VIROLI, C.: Mixtures of Quantile-based Factor Analyzers (pag. 126).

9:30 – 10:00 : Coffee break.

10:00 Session: Clustering, Classification and Discrimination 6 (A. Grané) (CC-DA2-1): `Room 1`.

10:00 – 10:20 PULIDO, B.: A multivariate approach for clustering functional data in one and multiple dimensions (pag. 122).

10:20 – 10:40 ORTEGO, M.: Analysis of seawater nutrient concentrations to assess Submarine Groundwater Discharge along the Catalan coast (NW Mediterranean): a Compositional Data Analysis Approach (pag. 116).

10:40 – 11:00 LABIOD, L. & NADIF, M.: Fuzzy Clustering of Attributed Networks (pag. 85).

10:00 Session: Data Science 3 (Social and Political Research) (DSc3-1): `Room 2`.

10:00 – 10:20 GUERRERO-SAN, M. & CUEVAS-COVARRUBIAS, C.: Crime in Mexico: an original Data Analysisapproach (pag. 78).

10:20 – 10:40 ARROYO-CASTRO, J. & CHOU-CHEN, S.: Unsupervised Detection of Anomaly in Public Procurement Processes (pag. 43).

10:40 – 11:00 BAKK, Z.: **CANCELLED** Unravelling the Skill Sets of Data Scientists: A Text Mining Analysis of Dutch University Master Programs in Data Science and Artificial Intelligence (pag. 46).

10:00 Session: LACSC session 4: Applications 2 (LACSC4-1): `Room 3`.

10:00 – 10:20 MARTÍNEZ, L. & , ROSTRÁN, A. &, BETANCO, G: Nicaragua Migrations: Origin-Destiniy (pag. 97).

10:20 – 10:40 MARADIAGA, E. & ROSTRÁN, A. & SOTO, A.: Traffic Accidents in Nicaragua (pag. 93).

11:10 Session: Plenary Talk 5 (Conf5): `Auditorium`.

11:10 – 12:10 MONTANARI, A.: (Data oblivious) Random Projections for (data aware) Model-based Clustering (pag. 108).

12:10 – 13:30 Time for lunch.

13:00 : Excursion. La Paz Waterfall Gardens

18:00 – : Conference Diner.

Thursday 18

08:30 Session: Model-Based Clustering 1 (M_BC1-1): `Room 1`.

08:30 – 08:50 QIN, X.: Clustering of human gut microbiome data using the finite mixture of generalized Dirichlet-multinomial models (pag. 123).

08:50 – 09:10 SCHARL, T.: A Clustering Procedure for Three-Way RNA Sequencing Data Using Data Transformations and Matrix-Variate Gaussian Mixture Models (pag. 133).

09:10 – 09:30 TAHIRI, T.: phyDBSCAN: phylogenetic tree density-based spatial clustering of applications with noise and automatically estimated hyperparameters (pag. 143).

08:30 Session: Data Science in Economics, Finance and Management (DScJ-1): `Room 2`.

08:30 – 08:50 NUÑEZ, M. & SCHNEIDER, M.: On the Vapnik-Chervonenkis Dimension and Learnability of the Hurwicz Decision Criterion (pag. 114).

08:50 – 09:10 MALECKA, M. & FISZEDER, P.: Robust estimation of the range-based GARCH model: Forecasting volatility, value at risk and expected shortfall of cryptocurrencies (pag. 91).

09:10 – 09:30 PIETRZYK, R. & MALECKA, M.: A Spectral Approach to Evaluating VaR Forecasts: Stock Market Evidence from the Subprime Mortgage Crisis, through COVID-19, to the Russo-Ukrainian War (pag. 120).

08:30 Session: Functional Data Analysis (FDA1-1): `Room 3`.

08:30 – 08:50 GÓRECKI, T. & KRZYŚKO, M. & WOLYŃSKI, W.: Applying classification methods for multivariate functional data (pag. 73).

08:50 – 09:10 MENDEZ, A.: A quantile extension to functional PCA (pag. 106).

09:10 – 09:30 VIDAL, M. & LEMAN, M. & AGUILERA, A.: Classification of neuroscientific data under the probabilistic principles of near-perfect classification (pag. 146). (ZOOM session)

09:30 – 10:00 : Coffee break.

10:00 Session: Machine Learning (ML1-1): `Room 1`.

10:00 – 10:20 MAKARENKOV, V.: Predicting soil bacterial and fungal communities at different taxonomic levels using machine learning (pag. 90).

10:20 – 10:40 PASQUIER, C. & SOLIS, M., VILCHEZ, V. & NÚÑEZ-CORRALES, S: Machine learning-driven COVID-19 early triage and large-scale testing strategies based on the 2021 Costa Rican Actualidades survey (pag. 118).

10:40 – 11:00 RAMIREZ, S. & GUEVARA VILLALOBOS, Á: Improving Employee Attrition with Data Analysis and Machine Learning (pag. 125).

10:00 Session: Data Science in Economics, Finance and Management 2 (DScJ1-1): `Room 2`.

10:00 – 10:20 QUIROS, T. & GUEVARA VILLALOBOS, Á: Innovating the banking with machine learning: Credit Score for MSMEs (pag. 124).

10:20 – 10:40 KUZIAK, K. & KACZMARCZYK, K. & COLAK, C.: Green bond yield determination with the use of machine learning methods. Comparison with conventional bonds (pag. 83).

10:40 – 11:00 JAJUGA, K.: Data Science in Finance classification of areas and methods (pag. 80).

10:00 Session: Functional Data Analysis 2 (FDA2-1): `Room 3`.

10:00 – 10:20 MATURO, F. & RICCIO, D. & ROMANO, E.: Improving Functional Classification Performance through Diversity: The Functional Voting Approach (pag. 104).

10:20 – 10:40 SCHILTZ, J. & NOEL, C.: Finite Mixture Models for an underlying Beta distribution with an application to COVID-19 data (pag. 134).

10:40 – 11:00 ANTON, C.: A multivariate functional data clustering method using parsimonious cluster weighted models (pag. 41).

11:10 – 11:40 Session: Award session: `Auditorium`.

- Chikio Ayasho Award
- Helga & Wolfgang Gaul Stiftung Aeard

11:40 Session: IFCS medal (IFCSmedal): `Auditorium`.

11:40 – 12:40 ROUSSEEUW, P.: New graphical displays for classification (pag. 128).

12:40 – 14:00 Time for lunch.

14:00 Session: Plenary Talk 7 (Conf7): `Auditorium`.

14:00 – 15:00 MÜLLER, A.: Similarity Search and LLMs: The RAG Revolution (pag. 110).

15:00 Session: Plenary Talk 6 (Conf6): `Auditorium`.

15:00 – 16:00 ALFARO, M.: Comparative study on Downscaling methods for Regional Climate Models (pag. 39).

16:00 – 17:00 : Poster session and Coffee break.

16:00 Session: Poster session (Poster): `Coffee area`.

- KRAL, C.: Model-based bi-clustering using multivariate Poisson-lognormal with general block-diagonal covariance matrix and its applications Submitted to IFCS 2024 Book of Abstracts (pag. 82).
  LIN, S.: On relation between separable effects, natural effects, and interventional effects
- MANNING, S.: Clustering for High-Dimensional, Nested Data with Categorical Outcomes Using a Generalized Linear Mixed Effects Model with Simultaneous Variable Selection (pag. 92).
- MARTINEZ, D. & MARTÍNEZ, K. & VELANDIA, D. & BUENDÍA, D: Spatial Agent-Based Model for Aedes Aegypti mosquitoes in the urban area in Arica (Chile) (pag. 96).
- PEREZ ALONSO, A. & ROSSEEL, Y. & VERMUNT, J.: Mixture Multigroup Structural Equation Modeling: Comparing Structural Relations Across Many Groups (pag. 119).
- SAME, A.: Drift-switching local level models for time series segmentation (pag. 132).

17:00 Session: Closing Plenary Talk (Conf8): `Auditorium`.

17:00 – 18:00 LUBBE, S.: A Collection of Biplots for Classification (pag. 89).

18:00 – 18:30 : Closing Ceremony

18:45 – ∞ : Closing Toast

Friday 19

08:30 Session: Tutorial 3 (Tut03): `Room 1`.

08:30 – 10:00 DE ROOIJ, M.: Logistic Multidimensional Data Analysis (pag. 67).

10:00 – 10:30 : Coffee break.

10:30 Session: Tutorial 3 (Tut03) (Cont...): `Room 1`.

10:30 – 12:00 DE ROOIJ, M.: Logistic Multidimensional Data Analysis (pag. 67).

12:00 – 13:30 Time for lunch.

13:30 Session: Tutorial 4 (Tut04): `Room 1`.

13:30 – 15:00 CHADJIPADELIS, T.: Methods on Artificial Intelligence (pag. 58).

15:00 – 15:30 : Coffee break.

15:30 Session: Tutorial 4 (Tut04)(Cont...): `Room 1`.

15:30 – 17:00 CHADJIPADELIS, T.: Methods on Artificial Intelligence (pag. 58).

6:45 – 22:00 : Farewell Dinner. La Casona de Lally (typical food), Curridabat

---

Contributions List[1]

[1] In strict alphabetic order according to the name of the author of the contribution.

Comparing Two Spatial Variables with the Probability of Agreement[I]

Communication

ACOSTA, JONATHAN[II] Vallejos, Ronny[III] Ellison, Aaron M.[IV] Osorio, Felipe[V]
de Castro, Mário[VI]

Chile

Computing the agreement between two continuous sequences is of great interest in statistics when comparing two instruments or one instrument with a gold standard. The probability of agreement (PA) quantifies the similarity between two variables of interest, and it is useful for determining what constitutes a practically important difference. In this article we introduce a generalization of the PA for the treatment of spatial variables. Our proposal makes the PA dependent on the spatial lag. We establish the conditions for which the PA decays as a function of the distance lag for isotropic stationary and nonstationary spatial processes. Estimation is addressed through a first-order approximation that guarantees the asymptotic normality of the sample version of the PA. The sensitivity of the PA with respect to the covariance parameters is studied for finite sample size. The new method is described and illustrated with real data involving autumnal changes in the green chromatic coordinate (22), an index of "greenness" that captures the phenological stage of tree leaves, is associated with carbon flux from ecosystems, and is estimated from repeated images of forest canopies.

Keywords: agreement, spatial variables, covariance.

[I] Tuesday 16, 08:50-09:10, Room 3, session: LACSC: Data Science and Computational Statistics: Theory and applications

[II] Pontificial Catholic University of Chile, Chile, jonathan.acosta@mat.uc.cl
[III] Technical University Federico Santa María, Chile, ronny.vallejos@usm.cl
[IV] Harvard University, United States, amellison17@gmail.com
[V] Technical University Federico Santa María, Chile, felipe.osorios@usm.cl
[VI] Universidade de São Paulo, São Carlos, Brazil, mcastro@icmc.usp.br

Combining Topic Modeling and Word Embedding to Predict Match Outcomes in Association Football[I]

Communication

ADHIKARI, SOURAV[II]

Austria

This study proposes a novel approach for predicting association football (soccer) match outcomes by leveraging text data from newspaper previews published in The Guardian, historical results and bookmakers' odds. Match results are categorized into a home team win, draw, or an away team win. Both Word2Vec and Llama 2 are used to extract features of the teams in form of word embedding. Topic modeling is subsequently applied to obtain match-specific features. Additionally, predictions from the Dixon and Coles model and pre-match bookmakers' odds are integrated to create an ensemble model. The feature set thus formed is utilized for a multi class classification using random forest classifier. Classification performance is assessed using several evaluation metrics such as precision, recall and F1 values after extensive cross validation. Upon comparison with existing approaches, the proposed method achieved an accuracy of 58.33% using text-based features and 64.5% with an ensemble approach. The presented methodology aims to provide an alternative way of extraction of team-specific and match-specific features along with enhancing prediction accuracy by integrating text and structured data.

Keywords: text analysis , topic modeling, large language models, feature engineering, classification.

[I]Tuesday 16, 17:40-18:00, Room 1, session: Data Mining

[II]Institute for Statistics and Mathematics, Vienna University of Economics and Business (WU Wien), Austria, sadhikar@wu.ac.at

Reproducible Data Analysis[I]

Tutorial

ALFARO, MARCELA[II]

United States

Reproducible data analysis is essential to ensure the transparency, collaboration, and validity of the results of any research project. Additionally, it is a strategy to avoid repeating internal processes in research or a private company and to document the various analyses that are periodically conducted in an organization. In this tutorial, we will describe the fundamental techniques to create and share reproducible data analyses using the R programming language, the version control package Git, and the collaboration platform GitHub, as well as how to use AI tools to document the code and make it more readable.

Introduction to Reproducible Data Analysis (Session 1: 30 minutes)

- What is reproducible data analysis and why is it important?
- Advantages and challenges of reproducibility in data analysis.
- Key tools and concepts: Quarto and version control.

Fundamentals of Quarto (Session 2: 1 hour)

- Integration of R code and narrative text.
- Structure and syntax of Quarto.
- Generation of dynamic and visually attractive reports.
- Incorporation of interactive graphs and tables.

Version Control with Git and GitHub (Session 3: 1 hour)

- Introduction to Git: tracking changes and collaboration.
- Creating and cloning repositories on GitHub.
- Integration of Quarto and Git for version tracking.
- Collaboration on data analysis projects.

Practice and Use of CoPilot Tools (Session 4: 30 minutes)

- Development of a short data analysis project from scratch.
- Using coPilot tools to document code.

Keywords: reproducible data analysis, R programming, version control (Git/GitHub).

[I] Monday 15, 08:30-12:00, Room 1, session: Tutorial 1

[II], United States,

Comparative Study on Downscaling Methods for Regional Climate Models[I]

Plenary Talk

ALFARO, MARCELA[II]

United States

---

egional climate models (RCMs) are essential for developing high-resolution climate outputs from general circulation models (GCMs). However, their computational demands are substantial, requiring significantly more compute time than statistical climate downscaling methods. This talk presents an innovative approach using a spatio-temporal statistical model with varying coefficients (VC) as a downscaling emulator for RCMs. We will describe the findings of our study, which compares two methods, INLA and varycoef, for estimating the proposed VC model. Through a comprehensive simulation setup, we evaluated the performance of these methods in constructing a statistical downscaling emulator for RCMs, with a particular focus on NARCCAP data. Our results show that the emulator effectively estimates non-stationary marginal effects, allowing for spatial variability in downscaling outputs. Additionally, the model demonstrates flexibility in estimating the mean of variables across space and time, achieving high predictive accuracy. INLA emerged as the fastest and most accurate method for parameter estimation and response variable distribution. This talk will also emphasize the importance of reproducible research practices for collaborative projects like this one, highlighting how transparency and verifiability can enhance scientific progress and collaboration.

Keywords: regional climate models (RCMs), general circulation models (GCMs), spatio-temporal statistical models, integrated nested Laplace approximations (INLA), NARCCAP, reproducible research.

[I]Thursday 18, 15:00-16:00, Auditorium, session: Plenary Talk 6

[II], United States,

Randomly Perturbed Random Forests[I]

Communication

ANDERLUCCI, LAURA[II] Montanari, Angela[III]

Italy

In supervised classification, a change in the distribution of a single feature, a combination of features, or the class boundaries, may be observed between the training and the test set. This situation is known as dataset shift. As a result, in real data applications, the common assumption that the training and testing data follow the same distribution is often violated.In order to address dataset shift we propose to randomly introduce more variability in the training set by sketching the input data matrix resorting to random projections of units. We then modify the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data.Results on real data show that perturbing the training data via matrix sketching improves the prediction accuracy of test units that have a different distribution in terms of variance structure.

Keywords: classification, dataset shift, data perturbation.

References

[1] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern recognition, 45(1), 521-530.

[2] Ahfock, D. C., Astle, W. J., & Richardson, S. (2021). Statistical properties of sketching algorithms. Biometrika, 108(2), 283-297

[3] Falcone, R., Anderlucci, L., & Montanari, A. (2022). Matrix sketching for supervised classification with imbalanced classes. Data Mining and Knowledge Discovery, 36(1), 174-208.

[I] Tuesday 16, 08:50-09:10, Room 1, session: Clustering, Classification and Discrimination 1

[II] University of Bologna, Department of Statistical Sciences, Italy, laura.anderlucci@unibo.it

[III] University of Bologna, Department of Statistical Sciences, Italy, angela.montanari@unibo.it

A Multivariate Functional Data Clustering Method Using Parsimonious Cluster Weighted Models[I]

Communication

ANTON, CRISTINA ADELA[II]

Canada

We propose a method for clustering multivariate functional linear regression data. Our approach extends multivariate cluster weighted models [2] to functional data with multivariate functional response and predictors, based on the ideas used by the funHDDC method [3]. To add model flexibility, we consider several two-component parsimonious models by combining the parsimonious models used for funHDDC with the Gaussian parsimonious clustering models family in [1]. Parameter estimation is carried out within the expectation maximization (EM) algorithm framework. The proposed method outperforms funHDDC on simulated and real-world data.

Keywords: cluster weighted models , functional linear regression, EM algorithm.

References

[1] Celeux G., Govaert G.: Gaussian parsimonious clustering models. Pattern Recognition **28**, 781–793 (1995)

[2] Dang U.J., Punzo A., McNicholas P.D., et al: Multivariate response and parsimony for Gaussian cluster-weighted models. J. Classif **34**, 4–34 (2017)

[3] Schmutz A., Jacques J., Bouveyron C., et al: Clustering multivariate functional data in group-specific functional subspaces. Comput. Stat. **35**, 1101–1131 (2020)

[I] Thursday 18, 10:40-11:00, Room 3, session: Functional Data Analysis 2

[II] MacEwan University, Department of Mathematics and Statistics, Canada, popescuc@macewan.ca

Principal Components Analysis of Histogram-valued Data: Set Theory Approach[I]

Communication

ARCE, JORGE[II]

Costa Rica

We introduce the Symbolic Principal Component Analysis method for Histogram-valued symbolic variables (HI-PCA), which represents an extension of the Symbolic Principal Component Analysis method for interval-valued symbolic variables (I-PCA). HI-PCA is based in classical set theory and probability theory, where we utilize the notions of bins and intervals to project histograms onto the principal components. The paper also presents several theorems that provide theoretical support for the HI-PCA method.We aim to explore the geometric aspects of projecting histogram bins onto Principal Components. The central question is whether, when applying PCA, if the intervals that support two bins are disjoint in the original data also they result in disjoint projections of the supporting intervals onto the principal components. Understanding these projections is crucial to us because they can impact the projected frequencies. To introduce HI-PCA based in set theory, we have developed definitions and theorems that offer a generalized approach based on the equations initially proposed in . Finally, we will illustrate the implementation of our proposed method in the R programming language using the RSDA package.

Keywords: symbolic data analysis, principal components analysis, histogram variables, bins, best point.

References

[1] Billard, L. and Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and DataMining (United Kingdom: John Wiley & Sons Ltd)

[2] Brito, P. and Dias, S. (2022). Analysis of Distributional Data CRC Press, United States of America.

[3] Rodriguez, O. (2023). RSDA: R to Symbolic Data Analysis. R package version 3.2.1

[4] Rodriguez, O. (2000). Classification et Modèles Linèaires en Analyse des Données Symboliques. Ph.D Thesis, Paris IX-Dauphine University.

[5] Verde, R., Irpino, A. and Balzanella A. (2016). Dimension Reduction Techniques for Distributional Symbolic Data. IEEE Transactions on Cybernetics 46(2)/ 344-355.

[I] Tuesday 16, 10:20-10:40, Room 2, session: Symbolic Data Analysis 2

[II] National University of Costa Rica, Costa Rica, jaag2486@gmail.com

Unsupervised Detection of Anomaly in Public Procurement Processes[I]

Communication

ARROYO-CASTRO, JOSE PABLO[II] Chou-Chen, Shu Wei[III]

Costa Rica

The procurement of goods and services in Public Administration is crucial for achieving institutional goals, with a focus on financial responsibility and trans- parent decision-making. In Costa Rica, public procurement is centralized through the Integrated Public Procurement System (SICOP). This study concentrates on goods procurement, aiming to identify successful contracts and detect anomalies. Machine Learning techniques, particularly under unsupervised approaches, enhance anomaly detection. The Principles of Integrity in Public Procurement Procedures from the Organisation for Economic Co-operation and Development (OECD) guide the evaluation process, emphasizing good procurement management, prevention of misconduct, and transparency. Various indicators, such as realistic budget estimation and objection rates, are utilized. Rapid procurement processes and price alterations may signal vulnerabilities and misconduct, highlighting the need for transparency and market awareness. Discovering its patterns is critical for accurate results, as different models respond differently to datasets and sample size changes. Emphasis should be placed on similar population clusters to avoid detecting natural anomalies. Implementing management mechanisms and employing data cleaning techniques are recommended to address data management errors.

Keywords: public procurement , machine learning, unsupervised learning, anomaly detection, corruption.

[I] Wednesday 17, 10:20-10:40, Room 2, session: Data Science 3 (Social and Political Research)

[II] Supreme Audit Institution of Costa Rica, Costa Rica, pabloarroyoc@gmail.com

[III] University of Costa Rica, Costa Rica, shuwei.chou@ucr.ac.cr

Pattern Recognition for Mexican Household Power Demand Time Series[I]

Communication

ASSE AMIGA, JOSÉ[II]

Mexico

Climate change is an issue caused in no small part by nonrenewableenergy generation [1] and low efficiency in electric systems [2]. A possible solution involves switching to renewable energy systems e.g. solar, wind, hydro, etc [2]. However, this presents a different problem that of intermittent power production [3]. Since the energy output cannot be controlled at any given time, a feasible strategy involves modifying consumption to fit into the supply, i.e. demand-side management (DSM). In order to use DSM strategies it is important to determine the possibility of demand-side flexibility, which is a measure of how demand can be modified to fit the supply [3]. To calculate flexibility it is necessary to first identify patterns within a power demand time series. This study presents a pattern recognition method developed for Mexican power demand time series, consisting of prepossessing, segmentation, dynamic time warping and clustering. Note that the data used is below the average Mexican household consumption rate of 4.76KWh [4], thus power demand can be volatile and present anomalies that can affect pattern segmentation [5]. The proposed solution involves using a triangular moving average to smooth the data. The next step is to separate the time series into possible patterns by analyzing rises and falls within the power for determined periods of time. Subsequently the distance between every sequence stored is calculated by using dynamic time warping, whichcan measure segments of different lengths in order to define relative closeness in shape [6]. By creating a distance matrix of every measure between two segments, a clustering algorithm can be applied, i.e. affinity propagation. This algorithm uses a similarity matrix as input and outputs the data grouped by closeness and defines centers for each group [7]. In the present use-case several patterns have been identified.

Keywords: pattern recognition, power time series, clustering.

References

[1] A.G. Olabi and Mohammad Ali Abdelkareem. Renewable energy and climate change. Renew.Sustain. Energy Rev., 158:112111, 2022.

[2] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points.Science, 315(5814):972–976, 2007

[3] International renewable energy agency. www.irena.org/Energy-transition/Outlook. Accessed:2024-04-02.

[4] Ivan Oropeza-Perez and Astrid H Petzold-Rodriguez. Analysis of the energy use in the mexicanresidential sector by using two approaches regarding the behavior of the occupants. Appl. Sci.,8(11):2136, 2018.

[I] Tuesday 16, 17:40-18:00, Room 2, session: Time Series Analysis and Pattern Recognition

[II] Instituto de Investigación Aplicada y TecnologíÂa, Universidad Iberoamericana, Mexico, joseasse55@gmail.com

[5] Nicole Nadine Ludwig. Data-Driven Methods for Demand-Side Flexibility in Energy Systems.PhD thesis, KIT, 2020.

[6] V. Tuzcu and S. Nas. Dynamic time warping as a novel tool in pattern recognition of ECGchanges in heart rhythm disturbances. In 2005 IEEE Trans. Syst. Man Cybern., volume 1, pages182–186. IEEE, 2005.

Unravelling the Skill Sets of Data Scientists: A Text Mining Analysis of Dutch University Master Programs in Data Science and Artificial Intelligence[I]

CANCELLED

BAKK, ZSUZSA[II]

The Netherlands

The growing demand for data scientists in the global labour market and the Netherlands has led to an increase in data science and artificial intelligence (AI) master programs offered by universities. However, there is still a lack of clarity regarding the specific skill sets of data scientists. This study aims to address this issue by employing Correlated Topic Modeling (CTM) to analyze the content of 41 master programs offered by 11 Dutch universities and an interuniversity combined program. We assess the differences and similarities in the core skills taught by these programs, determine the subject-specific and general nature of the skills, and provide a comparison between the different types of universities offering these programs. Our analysis reveals that data processing, statistics, research, and ethics are the core competencies in Dutch data science and AI master programs. General universities tend to focus on research skills, while technical universities lean more towards IT and electronics skills. Programs with a broad data science and AI focus generally concentrate on data processing, information technology, electronics, and research. In contrast, those with a subject-specific focus prioritize statistics and ethics. This research contributes to a better understanding of the diverse skill sets of Dutch data science graduates, providing insights for employers, academic institutions, and prospective students.

Keywords: data science, artificial intelligence (AI), skill sets, dutch universities, text mining.

CANCELLED

[I]Wednesday 17, 10:40-11:00, Room 2, session: Data Science 3 (Social and Political Research)

[II]Leiden university, The Netherlands, z.bakk@fsw.leidenuniv.nl

Statistics in the Knowledge Economy[I]

Communication

BANKS, DAVID[II]

United States

Statistics came of age when manufacturing was king. But todayÂ´s industries are focused on information technology. Remarkably, a lot of our expertise transfers directly. This talk will discuss statistics and AI in the context of computational advertising, autonomous vehicles, large language models, and process optimization

Keywords: autonomous vehicles, computational advertising, large language models.

[I]Tuesday 16, 14:50-15:10, Room 2, session: Advances in supervised classification

[II]Duke University, United States, banks@stat.duke.edu

Network Analysis Approach to the Analysis of Event Sequences[I]

Communication

BATAGELJ, VLADIMIR[II] Ferligoj, Aňuska, [III]

Slovenia

A person X's CV, $CV_X = (e_1, e_2, \ldots, e_{k_x})$, consists of a sequence of events e_i. For an event $e_i = (s_i, f_i, R_i, S_i, \ldots)$ we at least know its start date s_i, its end (finish) date f_i, the type R_i of the event, the state (location) S_i of the event, and maybe something more. We decided to base our analysis on the corresponding *co-presence network* – a weighted multi-relational temporal network $N = (V, L, w, t)$ in which the set of *nodes* V consists of studied persons. There is a *link* (edge) $\ell = (u : v$

Keywords: co-presence network, weighted multi-relational temporal network, temporal network.

[I]Tuesday 16, 14:50-15:10, Room 1, session: Data analysis

[II]University of Primorska, Slovenia, vladimir.batagelj@fmf.uni-lj.si

[III]University of Primorska,Slovenia, anuska.ferliogoj@fdv.uni-lj.si

Distributional-based Partitioning with Copulas[I]

Communication

BILLARD, LYNNE[II] Pan, Wenhao[III]

United States

An algorithm based on copula functions is considered for finding the partitions governing a data set consisting of a mixture of cumulative distribution functions.

Keywords: Archimedean copulas , elliptical copula, dynamical procedure.

[I]Tuesday 16, 08:30-08:50, Room 2, session: Symbolic Data Analysis 1

[II]University of Georgia, United States, lynne@stat.uga.edu

[III]Apple, United States, wenhao.pan@gmail.com

Robust Distance-based Generalized Linear Models: A New Tool for Classification[I]

Communication

BOJ, EVA[II] Grané, Aurea[III] Mayo-Íscar, Agustín[IV]

Spain

Understanding the nature of the data, dealing with outliers and redundant information are key issues when designing a proper metric for clustering and classification. Distance-based generalized linear models are prediction tools which can be applied to any kind of data whenever a distance measure can be computed among units. In this work, robust ad-hoc metrics are proposed to be used in the predictors' space of these models, incorporating more flexibility to this tool. Their performance is evaluated by means of a simulation study and compared to those based on Gower's and generalized Gower's metrics through several datasets of multivariate heterogeneous data with the presence of anomalous observations. Misclassification rate is used to evaluate the effectiveness in the prediction of responses. Additionally, ensemble methods are explored for such models in the context of big data. Applications on real data are provided in order to illustrate the predictive power of these models. Computations are made using the dbstats package for R.

Keywords: distance-based generalized linear models, robust metrics, ensemble methods.

[I] Tuesday 16, 17:00-17:20, Room 3, session: Classification Methods for Large Datasets (A. Grané)

[II] Universitat de Barcelona, Spain, evaboj@ub.edu

[III] Uniersidad Carlos III de Madrid, Spain, agrane@est-econ.uc3m.es

[IV] Universidad de Valladolid, Spain, agustin.mayo.iscar@uva.es

Scalable Conic Optimization for Feature Selection in Linear SVMs with Cardinality Control[I]

Communication

BOMZE, IMMANUEL M.[II] D'Onofrio, Federico[III] Peng, Bo[IV] Palagi, Laura [V]

Austria

In biclassification problems with many features, it is important to select a few of them, to ensure fairness and explainability in the original data space. We control the feature number (to be specified by the user) with a hard cardinality (zero-norm) constraint and solve the resulting difficult optimization problem by a conic decomposition approach. This exhibits promising scalability properties while maintaining both explainability and good predictive performance. From an optimization point of view, our approach is competitive with previous similar attempts employing mixed-binary linear optimization technology. These however use different techniques like bi-level formulations, surrogates for the zero-norm, or convex relaxationsIn biclassification problems with many features, it is important to select a few of them, to ensure fairness and explainability in the original data space. We control the feature number (to be specified by the user) with a hard cardinality (zero-norm) constraint and solve the resulting difficult optimization problem by a conic decomposition approach. This exhibits promising scalability properties while maintaining both explainability and good predictive performance. From an optimization point of view, our approach is competitive with previous similar attempts employing mixed-binary linear optimization technology. These however use different techniques like bi-level formulations, surrogates for the zero-norm, or convex relaxations, see for instance [1,2,3,4], and not all of them always exert strict control on the number of features., and not all of them always exert strict control on the number of features.

Keywords: biclassification, zero norm, conic optimization.

References

[1] Agor, Joseph and Özalt, Osman Y. (2019). Feature selection for classifcation models via bilevel optimization. Comput. OR 106, 156–1682.

[2] Aytug, Haldun (2015). Feature selection for support vector machines using Generalized Benders Decomposition. European J. OR 244, 210–2183.

[3] Ghaddar, Bissan and Naoum-Sawaya, Joe (2018). High dimensional data classifcation and feature selection using support vector machines. European J. OR 265, 993–10044.

[4] Labbé, Martine, Martínez-Merino, Luisa I. and Rodríguez-Chía, Antonio M. (2019). Mixed integer linear programming for feature selection in support vector machine. Discrete Appl.Math. 261, 276–304

[I] Tuesday 16, 15:10-15:30, Room 2, session: Advances in supervised classification

[II] University of Vienna, Austria, immanuel.bomze@univie.ac.at

[III] Sapienza Univ. of Roma, Italy, donofrio@diag.uniroma1.it

[IV] University of Vienna, Austria, bo.peng@univie.ac.at

[V] Sapienza Univ. of Rome, Italypalagi@diag.uniroma1.it

Candidates, Parties, Issues and the Political Marketing Strategies: A Comparative Analysis on political competition in Greece[I]

Communication

BOURANTA, VASILIKI[II] Panagiotidou, Georgia[III] Chadjipadelis, Theodore[IV]

Greece

This paper explores the evolving domain of political marketing, a field that extends beyond communication methods and public relations, encapsulating activities that influence the political behavior of parties and individual candidates. Drawing on theoretical frameworks and methodologies, we explore the application of marketing mix theory (product, price, place, promotion) within this political context. The focal point of our research is an in-depth examination of the political marketing strategy employed by Greek political parties during the Greek parliamentary elections of June 2023. The analysis scrutinizes the strategic patterns used in terms of selecting promotion tools, prioritizing political agenda issues, and focusing on the candidate versus the party. This involves advanced multivariate analysis methods such as Cluster Analysis, Multiple Correspondence Analysis, and Principal Component Analysis, which are utilized to detect and analyze in a comparative perspective the different strategies of the candidates and the parties in the Greek parliamentary elections of 2023. Moreover, the analysis focuses on how parties incorporated the newly implemented simple proportional representation system into their marketing strategies and their pre-electoral campaigns. Our data is derived from various sources including newspapers, mass media (TV, radio), and social media, allowing us to scrutinize the political product (party program and candidates), the 'price' (the voter's vote), the distribution strategies and promotion activities at both local and national level. Furthermore, we explore the relation between candidate profiles, their political marketing strategies, their political characteristics, and their probability of being elected or not. The paper suggests ultimately that political and electoral competition pivots on three pillars"candidates, parties, issues"which interact within the institutional framework as configured by the electoral law. This research bridges the gap between political marketing strategies, electoral systems, and their impact on campaign success, contributing significantly to the independent scientific scope of political marketing.

Keywords: political marketing , electoral campaign, greek elections, electoral systems, multivariate analysis.

[I] Wednesday 17, 08:50-09:10, Room 2, session: Data Science 1

[II] Aristotle University of Thessaloniki Greece, Greece, vickybouranta@gmail.com

[III] Aristotle University of Thessaloniki Greece, Greece, gvpanag@polsci.auth.gr

[IV] Aristotle University of Thessaloniki Greece, Greece, chadji@polsci.auth.gr

Quality Measures for Clusterwise Regression[I]

Communication

BRITO, PAULA[II] Dias, Sónia [III] Niang, Ndèye[IV]

Portugal

We focus on interval-valued variables, whose observations are intervals of real numbers. The Interval Distributional (ID) regression model [1] considers intervals represented by the corresponding quantile functions. The error between predicted and observed intervals, for each unit, is evaluated by the Mallows Distance. However, sometimes a single regression model is not appropriate, and it may be necessary to cluster the units and fit a regression model in each cluster. We apply a Clusterwise Regression model, for interval-valued variables, that finds the best partition of the data in clusters and simultaneously provides a linear regression model for each cluster. The algorithm [4], combines the dynamical clustering algorithm [2], and the ID regression model. The process is applied repeatedly varying the number of clusters K; for each fixed K, the algorithm considers different initial partitions and selects the solution with lowest Total Error. To select the best solution across different K, quality measures are proposed, that evaluate the fit between the clusters and their representing regression models. In particular, we extend the well-known Silhouette coefficient to clusterwise regression. The proposed model and measures are applied to a problem of pollution prediction in West Africa.

Keywords: clusterwise regression, interval data, silhouette coefficient.

References

[1] Dias, S. and Brito, P.: Off the beaten track: a new linear model for interval data. European Journal of Operational Research, 258(3), 1118–1130 (2017)

[2] Diday, E. and Simon, J.C.: Clustering analysis. In Digital Pattern Recognition, pp. 47–94. Springer (1976)

[3] Späth., H.: A fast algorithm for clusterwise linear regression. Computing, 29(2), 175–181 (1982)

[4] Suresh, N.: Clusterwise Linear Regression for Interval Data - An Extension of Interval Distributional Model. Master's thesis, Faculdade de Economia, Universidade do Porto (2020).

[I] Monday 15, 16:00-16:20, Room 2, session: Symbolic Data Analysis 3

[II] Fac. Economics, University of Porto & amp, Portugal, mpbrito@fep.up.pt

[III] ESTG, Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal, sdias@estg.ipvc.pt

[IV] Cédric-CNAM, Paris, France, ndeye.niang keita@cnam.fr

Nearest Neighbors for Mixed Type Data: An Inter–dependency Based Approach[I]

Chikio Hayashi Award

CAVICCHIA, CARLO[II] Iodice D'Enza, Alfonso[III] van de Velden, Michel[IV]

The Netherlands

K-Nearest Neighbors (KNN, [1]) is a relatively simple supervised method for both classification and regression tasks. It is referred to as a lazy-learner since the training observations are only used to identify the neighbors of the test observation. The prediction for each test observation is given by either neighborhood averaging (regression) or majority vote (classification). The neighbors identification is, therefore, the core of KNN and it critically depends on how the proximity (or, distance) between a test observation and a candidate neighbor is measured. There is a wide range of distance measures to choose from, specially for non continuous (categorical) data (see, e.g., [2]). In case of mixed type data, practitioners convert all the variable to a same type, or use ad-hoc combinations of distance measures, such as the Gower (dis)similarity index [3].We propose a KNN implementation that takes into account the inter-dependency structure among variables of mixed-type and consider two different approaches: association-based and entropy-based.

Keywords: nearest neighbors classification, mixed data.

References

[1] Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27

[2] Van De Velden, M., Iodice D'Enza, A. , Markos, A., and Cavicchia, C. (2023). A general framework for implementing distances for categorical variables. arXiv preprint arXiv:2301.02190.

[3] Gower, J. (1971). A general coefficient of similarity and some of its properties. Biometrics 27(4), 857–871

[I] Monday 15, 16:40-17:00, Room 3, session: Visualization (J. Nienkemper & S. Lubbe)

[II] Erasmus University Rotterdam, The Netherlands, cavicchia@ese.eur.nl

[III] University of Naples Federico II, Italy, iodicede@unina.it

[IV] Erasmus University Rotterdam, The Netherlands, vandevelden@ese.eur.nl

Predicting Air Pollution in Beijing, China Using Chemical, and Climate Variables[I]

Communication

CERVANTES ARTAVIA, JOSHUA ISAAC[II] Monge Cordonero, Moisés De Jesús[III]
Sabater Guzmán, Daniel Josué[IV]

Costa Rica

This study addresses atmospheric pollution, specifically in urban areas such as Beijing, China, focusing on PM2.5 particles. The importance of China in air pollution research and its correlation with meteorological factors and chemical compounds are emphasized. A forecasting model based on a state-space modeling approach is proposed to predict air pollution variation, utilizing data collected between 2013 and 2017 from various monitoring stations in Beijing. The theoretical analysis includes key concepts of air pollution, previous studies on PM2.5, as well as an introduction to time series analysis and state-space models. The results show that variables related to atmospheric pressure and wind speed are significant for predicting air pollution, although further exploration of additional methods for more precise variable selection is suggested. Furthermore, it is concluded that the proposed model is effective for short-term forecasts but may require refinement for longer periods.

Keywords: pollution , state-space model, time series.

[I] Tuesday 16, 15:10-15:30, Room 3, session: Applications

[II] Universidad de Costa Rica, Costa Rica, Joshua.cervantes@ucr.ac.cr
[III] Universidad de Costa Rica, Costa Rica, moises.mongecordonero@ucr.ac.cr
[IV] Universidad de Costa Rica, Costa Rica, daniel.sabater@ucr.ac.cr

Robustness under Missing Data: A Comparison with Special Attention to Inference[I]

Communication

CEVALLOS-VALDIVIEZO, HOLGER[II] Baum, Carole[III] van Messem, Arnout[IV]

Ecuador

The size of datasets is increasing at a rapid pace, both in terms of the number of observations as in the amount of included observed characteristics. Along with this, the probability that these datasets contain missing values rises as well. However, certain statistical processes and machine learning techniques are incapable of dealing with incomplete data. As such, it is of the utmost importance that these missing values are dealt with in an adequate way. Missing value imputation is a highly studied topic. A plethora of techniques have been proposed over the years to find suitable values to replace missing data, ranging from very simple techniques, such as mean or median imputation, to more complicated methods [2], such as the popular Multiple Imputation by Chained Equations method (MICE) [1]. With larger datasets, it is also more likely to observe a number of atypical or extreme data due to measurement and/or encoding errors. These outliers can, to a varying degree, influence statistical analyses. To alleviate this problem, robust techniques have been introduced.

Nowadays, imputation techniques are widely in use, but a large-scale comparison of these methods – and especially in terms of their robustness against outliers – seems to be missing. During a first attempt to fill this gap, we evaluate a large selection of imputation techniques, involving classic and robust procedures, by means of a simulation study with continuous data and different configurations of missing data and outliers. To evaluate the imputation capability and robustness of the imputation techniques we computed the mean prediction error between the actual data values and the predictions obtained by the imputation method. In this study, we also evaluated computational speed of the imputation methods. Our simulations indicate that, among the single imputation methods, robust linear regression using the MM-estimator and random forest imputation are among the most efficient and robust imputation methods, but these advantages naturally come at a cost, namely a higher computation time. However, often, the main concern is on the analysis that is performed after imputation. Therefore, in the second phase of our research, we evaluated the inferences and predictions made by different robust regression methods combined with an imputation technique in a simulation study with different configurations of outliers and missing data. For the simulations, we used a similar setting as in [3]. Both rowwise and cellwise outliers were generated, so we considered in the evaluation rowwise robust regression techniques as well as cellwise robust regression techniques. To evaluate the combined regression and imputation strategies in terms of inference capability, we measured the bias and variance of the estimated regression coefficients. To evaluate the prediction capability, we computed the mean prediction error.

Keywords: robustness, missing data, imputation, Inference.

[I] Tuesday 16, 10:20-10:40, Room 3, session: LACSC session 1: Data Science

[II] Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador, holgceva@espol.edu.ec

[III] University of Liège, Belgium, carole.baum@uliege.be

[IV] University of Liège, Belgiumarnout.vanmessem@uliege.be

References

[1] Stefan van Buuren and Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

[2] Wei-Chao Lin and Chih-Fong Tsai (2020). Missing value imputation: a review and analysis of the literature (2006-2017). *The Artificial Intelligence Review*, **53**(2), 1487–1509.

[3] Viktoria Öellerer, Andreas Alfons and Christophe Croux (2016). The shooting S-estimator for robust regression. *Computational Statistics*, **31**, 829–844.

Methods on Artificial Intelligence[I]

Tutorial

CHADJIPADELIS, THEODORE[II]

Greece

Using the example of an on-going research in the frame of AI4GOV project, we explore the connection between AI-driven digital transformation in municipalities and human rights, with a specific emphasis on Article 25(c) of the International Covenant on Civil and Political Rights, which advocates for equal access to public service, with the aim to evaluate how these technologies impact citizens' rights to equitable public service access. Public sector has increasingly integrated digital solutions to streamline services and enhance citizen engagement. However, aspects such as bias, inclusiveness and accessibility remain inadequately explored. Central to this investigation is the role of AI in either facilitating or hindering equal access to municipal services. The research employs a dual-method approach:

1. Qualitative Analysis: Conducting semi-structured interviews with municipal staff (internal users) to understand their experiences, challenges, and perceptions of the digital applications.

2. Quantitative Analysis: Survey to citizens (external end users) to assess the usability, accessibility, and inclusivity. The study aims to assess whether digital transformation initiatives are aligned with the principles of Article 25(c), ensuring that all citizens, irrespective of their background, have equitable access to these services. It also seeks to identify any digital barriers that might infringe upon this right, such as issues related to digital literacy, accessibility, and inclusivity.

By examining both internal and external perspectives on AI and digital applications in municipalities, this research is significant in the context of the evolving discourse on AI and human rights, offering a nuanced perspective on how digital transformation can both support and challenge the realization of fundamental human rights in the public sector. The findings are expected to contribute to the development of more inclusive and rights-aligned digital public services in Greek municipalities.

We investigate the impact of technology on the decision-making processes on a municipal level, and the interplay between civic engagement and emerging technologies. An informed and involved citizen is essential for the vitality of democratic governance, and European countries work to ensure the active participation of individuals in political life, and their engagement in the decision-making processes. Simultaneously, local governance acknowledges the role of technology as an enabler of civic participation, with digital platforms and e-governance initiatives serving as tools to amplify voices and streamline decision-making. We explore the dynamic relationship between technology-driven civic participation and the evolution of decision-making mechanisms at the municipal level.

Municipalities today aim for the integration of state-of-the-art technological solutions such as Artificial Intelligence (AI) and Blockchain technologies, establishing a transparent and secure platform that redefines citizen involvement in decision-making. Through the utilization of such emerging technologies, citizens are

[I] Friday 19, 13:30-17:00, Room 1, session: Tutorial 4

[II] Aristotle University of Thessaloniki, Greece, chadji@polsci.auth.gr <chadji@polsci.auth.gr>

equipped with advanced and secure tools to express their opinions on matters affecting the community, fostering a sense of ownership that can lead to a more politically informed and engaged society.

Drawing experience and input form the AI4Gov project, we present the utilization and integration of AI and blockchain technologies on the municipal level to optimize resource management, reduce environmental impact, and elevate overall quality of life. Through AI-based solutions, such as smart waste management systems, traffic violation solutions, and civic engagement tools, the municipality enhances accessibility for all residents, irrespective of age or ability.

The integration of blockchain ensures the security, audit, integrity and transparency of the data being analyzed. It enables these technologies to play vital roles in reinforcing and upholding democratic principles and strengthening democratic processes and governance worldwide. Through civic participation, Europe builds resilient and responsive democracies that reflect the diverse perspectives and needs of its citizens.

Contents

Keywords: articial intelligence (AI), qualitative methods, quantitative methods, correspondence analysis, semantic map, semantic analysis.

Towards Topologically Diverse Probabilistic Planning Benchmarks: Synthetic Domain Generation for Markov Decision Processes[I]

Contributed Sessions

CHAMPAGNE GAREAU, JAËL[II] Beaudry, Éric[III] Makarenkov, Vladimir[IV]

Canada

Markov Decision Processes (MDPs) are often used in Artificial Intelligence (AI) to solve probabilistic sequential decision-making problems. In the last decades, many probabilistic planning algorithms have been developed to solve MDPs. However, the lack of standardized benchmarks makes it difficult to compare the performance of these algorithms in different contexts. In this paper, we identify important topological properties of MDPs that can make a significant impact on the relative performance of probabilistic planning algorithms. We also propose a new approach to generate synthetic MDP domains having different topological properties. This approach relies on the connection between MDPs and graphs and allows every graph generation technique to be used to generate synthetic MDP domains.

Keywords: markov decision process, probabilistic planning, synthetic domains generation, topological diversity, benchmarking.

[I] Monday 15, 17:25-17:45, Room 2, session: Optimization in Classification and Clustering

[II] Université du Québec à Montréal, Canada, champagne_gareau.jael@uqam.ca

[III] Université du Québec à Montréal, Canada, beaudry.eric@uqam.ca

[IV] Université du Québec à Montréal, Canada, makarenkov.vladimir@uqam.ca

Symbolic Data Analysis Framework for Recommendation Systems: SDA-RecSys[I]

Communication

CHAPARALA, PUSHYA[II]

India

Recommendation algorithms, often rely on user-item interaction matrices, to uncover hidden patterns and preferences. These matrices play a pivotal role in facilitating the detection of matching similarities between users and items. However, these matrices do not capture the full spectrum of users' preferences in ratings while providing a list of recommendations. Since such variability can be effectively modeled as symbolic objects, specifically histogram objects, it is proposed to use the Symbolic Data Analysis (SDA) tools to address this challenge. This inclusion of user preferences and item characteristics into histograms enhanced the user profile capabilities in our methodology. These profiles can then be compared using Wasserstein similarity measures to compute the nearness between users and items, enabling the recommender system to generate top-N relevant recommendations. To evaluate the efficacy of the proposed SDA-RecSys, experiments are conducted to assess the impact of histogram profiles on recommendations, by utilizing the Normalized Discounted Cumulative Gain (NDCG) metric as a benchmark. Comparisons are presented to project the superiority of the SDA framework for Recommendation systems.

Keywords: information overload , recommender systems, histogram objects, symbolic data analysis (SDA).

[I]Monday 15, 16:20-16:40, Room 2, session: Symbolic Data Analysis 3

[II]Department of Computer Science and Engineering Vignan's Foundation for Science, Technology , India, pushyachaparala@gmail.com

Bayesian Selection Approach for Categorical Responses via Multinomial Probit Models[I]

Plenary Talk

CHEN, RAY-BING[II]

Taiwan

A multinomial probit model is proposed to examine a categorical response variable, with the main objective being the identification of influential variables in the model. To this end, a Bayesian selection technique is employed, featuring two hierarchical indicators where the first indicator denotes a variable's relevance to the categorical response, and the subsequent indicator relates to the variable's importance at a specific categorical level, aiding in assessing its impact at that level. The selection process relies on posterior indicator samples generated through an MCMC algorithm. The efficacy of our Bayesian selection strategy is demonstrated through both simulation and an application to a real-world example.

Keywords: indicator, MCMC Algorithm, multi-task learning.

[I]Tuesday 16, 11:10-12:10, Auditorium, session: Plenary Talk 2

[II]National Cheng Kung University, Taiwan, rbchen@ncku.edu.tw

Integration of Data from Probability and Non-probability Surveys[I]

Plenary Talk

COBO, BEATRIZ[II]

Spain

Over the years, probability surveys have been the gold-standard since statistical developments are based on them. Over the years, non-probability surveys have been used more and more and they saw their rise in the pandemic period, since it was the only way to collect data, but they have the problem that they do not have a defined sampling frame. The idea is to take the best of both, for example, some authors use non-probability samples to obtain new weights with which to estimate the variables of interest more efficiently. In this case, we are going to combine both surveys, weighting the estimates in different ways with the aim of obtaining reliable estimates with less error.

Keywords: probability surveys, non-probability surveys, data weighting techniques.

[I] Tuesday 16, 16:00-17:00, Auditorium, session: Plenary Talk 4

[II] Department of Quantitative Methods for Economics and Business, University of Granada, Spain, beacr@ugr.es

Gender Bias Mitigation in a Credit Scoring Model[I]

Communication

CORRALES-BARQUERO, RICARDO[II]

Costa Rica

A study carried out on a dataset and a mathematical model to support decision making in the credit process for established clients in a commercial bank in Costa Rica will be presented. The main objective of the study was to evaluate alternatives to mitigate the gender biases present in the model. To achieve this,possible sources of bias in the modelwere identified, among which possible disparate treatment, association, selection, malicious, and automation biases were identified.These biases were then measured in more detail, finding that they are small, except perhaps for the selection bias. Thirdly, alternative models were built to mitigate these biases, to finally evaluate the difference both in the fairness measures that were used and in the performance of the alternative models compared to the original to determine the one that provides greater value to the business. Here, it was found that the gains are minor and that what could be more worthwhile is to maintain the current model and investigate other credit scoring models used in other stages of the credit granting process.

Keywords: gender bias, bias mitigation, credit scoring.

[I] Wednesday 17, 09:10-09:30, Room 2, session: Data Science 1

[II] Banco Nacional de Costa Rica, Dirección de Modelos Matemáticos Universidad de Costa Rica, Programa de Posgrado en Computación e Informática, Costa Rica, ricorrales07@gmail.com

A Deterministic Information Bottleneck Method for Clustering Mixed-Type Data[I]

Communication

COSTA, EFTHYMIOS[II] Papatsouma, Ioanna[III] Markos, Angelos[IV]

Great Britain

In this paper, we present an information-theoretic method for clustering mixed-type data, that is, data consisting of both continuous and categorical variables. The method is a variant of the Deterministic Information Bottleneck algorithm which optimally compresses the data while retaining relevant information about the underlying structure. We compare the performance of the proposed method to that of three well-established clustering methods (KAMILA, K-Prototypes, and Partitioning Around Medoids with Gower's dissimilarity) on simulated and real-world datasets. The results demonstrate that the proposed approach represents a competitive alternative to conventional clustering techniques, particularly in scenarios with unbalanced clusters and significant overlap between clusters.

Keywords: deterministic information bottleneck , clustering, mixed-type data, mutual information.

[I] Tuesday 16, 08:30-08:50, Room 1, session: Clustering, Classification and Discrimination 1

[II] Imperial College London, Department of Mathematics, Great Britain, efthymios.costa17@imperial.ac.uk

[III] Imperial College London, Department of Mathematics, Great Britain, i.papatsouma@imperial.ac.uk

[IV] Democritus University of Thrace, Greece, amarkos@eled.duth.gr

New Tools for Visualizing Distributional Datasets[I]

Communication

CROCETTA, CORRADO[II] Irpino, Antonio[III]

Italy

Visualization tools are crucial for conveying patterns and guiding analytical approaches in data interpretation. When faced with the complexity of visualizing distributional data tables, where each observation is a vector of frequency or density distributions, the need for user–friendly tools becomes apparent. To address this challenge, three innovative visualization tools have been introduced for data tables characterized by numeric symbolic distributional data.The first two tools, the Green Eye Iris (GEI) and the Flower plot, utilize a polar coordinate-based representation of stacked bar charts or violin plots. The third tool extends the traditional heatmap plot and is particularly effective for illustrating datasets with numerous observations and variables. All three methods focus on visually representing the proportion of mass distributed on the domain variable, utilizing diverging color palettes for each distribution.

Keywords: symbolic data, distributional data, visualization.

[I] Monday 15, 17:00-17:20, Room 2, session: Symbolic Data Analysis 3

[II] Italian Statistical Society, University of Bari, Italy, corrado.crocetta@uniba.it

[III] University of Campania L. Vanvitelli, Italy, antonio.irpino@unicampania.it

Logistic Multidimensional Data Analysis[I]

Tutorial

DE ROOIJ, MARK[II]

The Netherlands

This tutorial focuses on the analysis of multivariate categorical response variables. For categorical variables, logistic models are most usual. However, having more than one response variable the data and its analysis become complex. In this tutorial, a logistic analysis framework based on dimensionality reduction techniques like principal component analysis, reduced rank regression, and (restricted) multidimensional unfolding for the analysis of multivariate categorical variables is presented. In this framework, the negative log-likelihood is minimized for which a Majorization Minimization (MM) algorithm is used. Special attention is given to biplots for the graphical representation of the results. Theory and applications will be shown in detail. Also, R–software for the analysis of data will be shown.

Content

- Logistic regression models for binary, ordinal, and nominal variables
- Principal component analysis, reduced rank regression, and (restricted) multidimensional unfolding
- Logistic Multidimensional Data Analysis for binary variables
- Logistic Multidimensional Data Analysis for ordinal variables
- Logistic Multidimensional Data Analysis for nominal variables

Keywords: biplots, classification, dimension reduction, logistic regression, multi-label.

[I]Friday 19, 08:30-12:00, Room 1, session: Tutorial 3

[II]Leiden University, The Netherlands, rooijm@fsw.leidenuniv.nl

Reduced Rank Regression with Mixed Predictors and Mixed Responses[I]

Communication

DE ROOIJ, MARK[II]

The Netherlands

We propose generalized mixed reduced rank regression for the analysis of mixed response variables and mixed predictor variables. The response variables can be a mixture of numeric, ordinal, and binary variables for which we combine, in a single model, ideas from linear regression, logistic regression, and cumulative logistic regression. The predictor variables can be a mix of binary, nominal, ordinal, and numeric data. For categorical predictor variables, we propose to use optimal scaling, that provides optimal quantifications of the predictor variables. All these elements are combined into a single multivariate regression model, where we place a rank restriction on the matrix with regression coefficients to reduce the dimensionality. A majorization-minimization algorithm is proposed for maximum likelihood estimation of the model parameters. The methodology will be illustrated using data from the Eurobarometer survey.

Keywords: MM algorithm, nominal, ordinal, numeric, dichotomous.

[I] Monday 15, 16:00-16:20, Room 3, session: Visualization (J. Nienkemper & S. Lubbe)

[II] Leiden University, The Netherlands, rooijm@fsw.leidenuniv.nl

Multimodal Emotion Recognition: A Comparative Study[I]

Communication

FALIH, ISSAM[II] Gondech, Ayemn[III] Tran, Hélène[IV]

France

Recent advancements in machine learning have highlighted the importance of integrating different data sources to improve classification model performance. By utilizing multiple data representations, a richer understanding of subjects or objects can be achieved. For instance, in emotion recognition field combining multiple sources and /or modalities of information (e.g., voice, text, facial expression, body posture) performs well than those relying solely on a single modality. The challenge lies in fusing distinct types of data such as image, text, audio or video that are not naturally aligned.Traditional classification algorithms, initially designed for uni-modal datasets, struggle with the complexities presented by multi-modal scenarios. This complexity is exacerbated by the need to align heterogeneous data sources, manage increased dimensionality, and create complementary and non-redundant representations. To tackle these issues, two principal family of approaches have emerged. The first is `agnostic` to the specific model, focusing on the moment when the fusion occurs along with the nature of the fusion, i.e: early (feature-level), late (decision-level) or hybrid. The second family of approaches is `model-dependent`, which involves sophisticated techniques like kernel methods, graphical methods, and deep neural networks. These strategies aim to consider the full potential of multi-modal data, thereby significantly elevating the capabilities of classification models. For each family of approaches, different techniques exist and in this work we aim to highlight the main methods tackling this problem and we present a comparative study on different multi-modal datasets. Additionally, we outline prospective directions for future research in this evolving field.

Keywords: multimodality, classification, deep learning, emotion recognition.

[I] Tuesday 16, 10:00-10:20, Room 1, session: Clustering, Classification and Discrimination 2

[II] LIMOS, Clermont Auvergne University, France, issam.falih@uca.fr

[III] LIMOS, Clermont Auvergne University, France, Aymen.GONDECH@etu.uca.fr

[IV] LIMOS, Clermont Auvergne University, France, Helene.TRAN@doctorant.uca.fr

Comparison of Cooling Schedules in the Simulated Annealing Algorithm Applied to the Clustering Problem[I]

Communication

FALLAS MONGE, JUAN JOSÉ[II]

Costa Rica

The simulated annealing (SA) algorithm was proposed by Kirkpatrick, Gellat, and Vecchi (1983). This algorithm evaluates new solutions based on a cooling rate known as *temperature*, denoted by T_k. Two widely accepted models for defining T_k exist. The first is the geometric model: $T_k = \alpha T_{k-1}$, with $0 < \alpha < 1$, which is most commonly used in practice. The second is the logarithmic model: $T_{k+1} = \frac{c}{\log(k+1)}$,where c is a constant, facilitating rigorous convergence tests of the SA algorithm. While these are prevalent, there are other cooling schedules less commonly employed yet utilized in research applying the simulated annealing algorithm. Examples include:

$T_{k+1} = T_k \cdot \left\{1 + \frac{T_k \cdot \ln(1+\delta)}{3 \cdot \sigma_k}\right\}^{-1}$, $T_{k+1} = \frac{T_k}{1+\beta \cdot T_k}$, $T_{k+1} = \frac{T_k}{1+\beta_k \cdot T_k}$

and

$T_k = T_0 \cdot a^{-\left[\frac{k}{f \cdot k_x}\right]^b}$.

In this presentation, we will compare these cooling schedules to assess whether the widespread adoption of the geometric model is truly justified. This comparison will be conducted within the context of an optimization problem associated with partitioning a set of quantitative data.

Keywords: simulated annealing, cooling schedules, clustering.

[I] Tuesday 16, 10:40-11:00, Room 3, session: LACSC session 1: Data Science

[II] Instituto Tecnológico de Costa Rica, Cartago, Costa Rica, jfallas@itcr.ac.cr

Visualization and Clustering with Projective Techniques[I]

Communication

FRANCE, STEPHEN L.[II]

United States

Projective techniques have their origins in personality research and have been utilized in marketing and consumer research for over fifty years. They are used to elicit deeper understanding of a brand, service, or other business entity. Projective techniques typically utilize questions that link the concept being studied to specific tasks, objects, or other people. For example, for a retailer, one may ask "This retailer remains you of which animal?" or "If the retailer is a relative then which relative would this be?" Projective techniques come under the banner of qualitative research, but the data can be quantified and analyzed using quantitative methods such as multidimensional scaling (MDS) and cluster analysis. This talk looks at methods for exploring structure using projective techniques. A specific focus is given on how to optimally pre-process the discretized answer information. This can be structured using a bag-of-words approach, but with very uneven density. For example, when comparing a brand to an animal, some animals, such as lions, are used very often, which other animals, for example, walruses, are rarely used. This talk will describe methods of optimally weighting observations to maximize the information obtained relative to error and to improve the accuracy of brand mapping visualizations and clusterings . An example will show how a range of well known US retail brands are viewed and will give insights to consumer perceptions of brands relative to projective traits, including similarities with animals, cars, music artists, and American cities. An extension method is described where the attributes of the projective answers are crowdsourced from the web and used to improve the accuracy of the analysis.

Keywords: projective techniques , cluster analysis, MDS, marketing.

[I] Tuesday 16, 17:20-17:40, Room 1, session: Data Mining

[II] Mississippi State University Marketing, Quantitative Analysis, and Business Law, United States, sfrance@business.msstate.edu

Classifying Multivariate Observations in Data Sets with Asymmetric Features and Outlying Observations[I]

Communication

FRANCZAK, BRIAN[II]

Canada

Classification can be defined as the process of sorting similar objects into groups. Classification can be performed in unsupervised, semi-supervised, or fully supervised settings. In the unsupervised setting, also known as clustering, no prior information is used, while the other two settings use some prior knowledge. Model-based clustering is the process of using a finite mixture model for unsupervised classification. This talk will discuss an approach for performing model-based clustering and outlier detection for incomplete multivariate data sets. A expectation-maximization (EM) based parameter estimation scheme is discussed and utilized for the considered mixtures of contaminated shifted asymmetric Laplace distributions. This EM based scheme iteratively performs single imputation while estimating the maximum likelihood estimates of the model of interest. At convergence, we use traditional likelihood-based criteria like the Bayesian information criterion for model selection. We assess classification performance using the adjusted Rand index and give other relevant statistics demonstrating the overall performance of the parameter estimation scheme. We demonstrate the effectiveness of the proposed model using simulated and real data sets.

Keywords: model-based clustering, outlier detection, imputation, finite mixture models, expectation-maximization algorithm.

[I]Monday 15, 16:20-16:40, Room 1, session: Modeling Multivariate Data (A. Roy)

[II]MacEwan University, Canada, franczakb@macewan.ca

Applying Classification Methods for Multivariate Functional Data[I]

Communication

GÓRECKI, TOMASZ[II] Krzyśko, Mirosław[III] Wołyński, Waldemar[IV]

Poland

In this article, we propose a new approach to the classification of multivariate time series. We use a functional approach to data analysis and combine information from raw data and functional derivatives. To provide a comprehensive comparison, we conducted a set of experiments, testing effectiveness on fifteen multivariate time series datasets from a wide variety of application domains. Our experiments show that this new method provides a more accurate classification of the examined datasets.

Keywords: functional data , classification, discriminant coordinates, curvature.

[I] Thursday 18, 08:30-08:50, Room 3, session: Functional Data Analysis

[II] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, 61-614 Poznań, Poland, Poland, drizzt@amu.edu.pl

[III] University of Kalisz, 62-800 Kalisz, Poland, Poland, mkrzysko@amu.edu.pl

[IV] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, 61-614 Poznań, Poland, Poland, wolynski@amu.edu.pl

High-dimensional Survival Analysis: Exploring Cox Regression with Lasso and Adaptive Lasso Penalties[I]

Communication

GONZÁLEZ-BARQUERO, PILAR[II]

Spain

This work is centered on the application of survival analysis to a high-dimensional dataset provided by the Gregorio Marañón Health Research Institute. The dataset contains clinical and genetic information from patients with triple-negative breast cancer (TNBC), a type of cancer known for its aggressive nature and low survival rates. The patients in the dataset have been treated with a specific type of chemotherapy, and their survival time is measured from the beginning of the treatment until death.The main objective of this work is to classify variables (genetic and clinical) based on their influence on the survival of these patients. To achieve this, we assess the effectiveness of Cox regression models (Cox, 1972) in the context of high-dimensional data and high proportion of censure. Dimensionality reduction techniques are crucial for model interpretability and predictive accuracy in this context. Two regularization techniques, the lasso penalty (Tibshirani, 1997) and the adaptive lasso penalty (Zou, 2006), are evaluated. The main contributions of this work are the proposal of different adaptive weight calculation methods for the adaptive lasso, and a new procedure for finding the best model or variable selection for Cox regression.

Keywords: survival analysis , Cox regression, high-dimensional data.

[I] Tuesday 16, 17:20-17:40, Room 3, session: Classification Methods for Large Datasets (A. Grané)

[II] University Carlos III of Madrid, Spain, mariapilar.gonzalez@uc3m.es

A New Distance for Categorical Data with Moderate Association[I]

Communication

GRANÉ, AUREA[II] Salini, Silvia[III] Infante, Gabriele[IV]

Spain

Categorical variables coming form surveys usually share high percentages of information. Redundant information may lead to misleading results in data visualization techniques and clustering procedures, since units with similar characteristics can be considered as completely different. In general, this situation is encountered when additive dissimilarity coefficients are used in datasets with moderate or high association, producing the typical horseshoe effect which arises when visualizing the data in low-dimension. In this work we propose a new distance for categorical data, able to take into account the association/correlation structure of the data. Its performance is evaluated and compared to Hamming distance in MDS configurations. Additionally, applications to novel data on co-creation antecedents of telemedicine and vehicle accident rate data are given to illustrate the methodology.

Keywords: association , categorical data, horseshoe effect, MDS, redundant information.

[I] Wednesday 17, 08:30-08:50, Room 1, session: Clustering, Classification and Discrimination 5 (A. Grané)

[II] Universidad Carlos III de Madrid, Spain, aurea.grane@uc3m.es

[III] University of Milan, Italy, silvia.salini@unimi.it

[IV] University of Milan, Italy, gabriele.infante@unimi.it

MM-Algorithms in Data-Science[I]

Plenary Talk

GROENEN, PATRICK[II]

The Netherlands

As an optimization method, majorization and minorization (MM) algorithms have been applied with success in a variety of models arising in the area of statistics and data science. A key property of majorization algorithms is guaranteed descent, that is, the function value decreases in each step. In practical cases, the function is decreased until it has converged to a local minimum. If the function is convex and coercive, a global minimum is guaranteed. The auxiliary function, the so-called majorizing function, is often quadratic so that an update can be obtained in one step. Here, we present a selection of useful applications of MM algorithms. We discuss its use multidimensional scaling, and in binary and multiclass classification such as logistic regression, multinomial regression, and support vector machines. In the case of regularized generalized canonical correlation analysis, its MM algorithm coincides with several partial least squares (PLS) algorithms thereby providing a previously unknown goal function for the PLS algorithms. We show how MM can also be effective in large scale optimization problems, such as the SoftImpute approach for dealing with missings in principal components analysis, and the MM algorithm for convex clustering.

As an optimization method, majorization and minorization (MM) algorithms have been applied with success in a variety of models arising in the area of statistics and data science. A key property of majorization algorithms is guaranteed descent, that is, the function value decreases in each step. In practical cases, the function is decreased until it has converged to a local minimum. If the function is convex and coercive, a global minimum is guaranteed. The auxiliary function, the so-called majorizing function, is often quadratic so that an update can be obtained in one step. Here, we present a selection of useful applications of MM algorithms. We discuss its use multidimensional scaling, and in binary and multiclass classification such as logistic regression, multinomial regression, and support vector machines. In the case of regularized generalized canonical correlation analysis, its MM algorithm coincides with several partial least squares (PLS) algorithms thereby providing a previously unknown goal function for the PLS algorithms. We show how MM can also be effective in large scale optimization problems, such as the SoftImpute approach for dealing with missings in principal components analysis, and the MM algorithm for convex clustering.

Keywords: majorization and minorization (mm) algorithms , multidimensional scaling, regularized generalized canonical correlation analysis.

[I] Monday 15, 14:30-15:30, Auditorium, session: Opening Plenary Talk

[II] Erasmus University, Rotterdam, Econometric Institute, The Netherlands, groenen@ese.eur.nl

Optimal Penalized Sparse PCA[I]

Communication

GUERRA URZOLA, ROSEMBER ISIDORO[II]

The Netherlands

In Machine Learning and Data Science, there is a growing interest in algorithms that yield sparse solutions due to their interpretability and computational efficiency. Penalized methods are commonly employed to achieve sparsity in Principal Component Analysis (PCA) problems. However, a key critique of these methods is that their performance is assessed via numerical experiments, lacking theoretical assurances of optimality. Our research aims to explore the theoretical properties that determine the success of one of these algorithms in achieving optimality.Specifically, we concentrate on penalized PCA formulations utilizing cardinality as a sparsity-inducing penalty. We introduce a Minorization-Maximization scheme to tackle the problem and theoretically demonstrate that the resulting solution constitutes a local optimum. We establish essential conditions for optimality. We require the minimum eigenvalue of the sample covariance matrix greater than one, leading to the absence of a feasible ascent direction at the solution. However, this condition may not be applicable in practice, especially in high-dimensional scenarios.To address this challenge, we propose a straightforward procedure to ensure adherence to this condition across diverse datasets. Subsequently, we conduct numerical experiments utilizing both synthetic and empirical datasets to elucidate the practical implications of this condition.

Keywords: dimension reduction, sparse solutions, principal component analysis.

[I]Monday 15, 17:25-17:25, Room 1, session: Dimension Reduction

[II]Tilburg University, Department of methodology and statistics, The Netherlands, r.i.guerraurzola@tilburguniversity.edu

Crime in Mexico: An Original Data Analysis Approach[I]

Communication

GUERRERO-SAN VICENTE, MARÍA TERESA[II] Cuevas-Covarrubias, Carlos[III]

Mexico

Crime and violence are the main social problem in Mexico. Insecurityaffects important aspects of daily life: family, social relations, economy, education,work, health and leisure activities. We present a statistical analysis of different lifestyle aspects of young men in Mexico. This variables were previously reported bythe government in ECOPRED 2014, a survey of national coverage performed by theNational Institute of Statistics. The main contribution of this paper is a revealing dataanalysis based on a combination of linear discriminant scores and Mutual PrincipalComponents where the area under the ROC curve works as the link between boththechniques. It describes the relation between the life style of young men and their riskof incurring in criminal behavior. This risk is measured in terms of a uni dimensionallinear score with high discrimination capacity.

Keywords: delincuency, antisocial behavior, discrimination, mutual PCA, ROC curves.

References

[1] Cuevas-Covarrubias,C. : Principal Components Analysis for a Gaussian Mixture. In: Lausen,B., Van den Poel, D., Ultsch, A. (eds.) Algorithms from and for Nature and Life, pp. 175-183.Springer, Heidelberg (2013)

[2] Guerrero-San Vicente, M.T., Cuevas-Covarrubias, C. : Conductas de riesgo en jóvenes mexicanos,detección de factores de riesgo y modelación estadística. In: Martínez Lánz, P. (ed.)Delincuencia en México, pp. 53-71. Editorial Porrua, México (2022).

[I] Wednesday 17, 10:00-10:20, Room 2, session: Data Science 3 (Social and Political Research)

[II] Universidad Anáhuac, México, Mexico, teresa.guerrero@anahuac.mx

[III] Universidad Anáhuac, México, Mexico, ccuevas@anahuac.mx

Analysis of Intoxication Cases Reported in Costa Rica from 2020 to 2022: Before and During the COVID-19 Period.[I]

Communication

GUTIERREZ VEGA, EDGARDO[II] Chou-Chen, Shu-Wei[III]
Somarribas-Blanco, Marietta[IV]

Costa Rica

The pandemic caused by SARS-CoV-2 and the health restrictions led to drastic changes in the lifestyles and habits of the population. The main goal of this work is to determine significant changes during the COVID-19 period (2020 to 2022) compared to previous years (2015 to 2019), by using reports received by the National Intoxication Control Center of Costa Rica. First, a hierarchical clustering analysis using Dynamic Time Warping (DTW) was performed to identify relevant time series based on sex and age groups for each intoxication cause. Then, we focused on these time series for each intoxication cause and analyzed if they suffered changes when the pandemic started. A forecasting approach using ARIMA with intervention and Prophet models was used to determine unexpected behavior after the start of the pandemic. The results showed that children 0 to 4 years old were mainly affected by accidental intoxications, with a decreasing trend during the pandemic.Regarding attempted suicide with medication, two main groups were analyzed: women aged 12 to 19 years old and adults aged 20 to 59 years old, both showing an increase in cases. Reports of addiction with drugs of abuse were analyzed for individuals aged 12 to 19 years old and those over 20 years old, with only the latter showing an increase. For reactions to medications, individuals under 15 years of age presented a decrease in cases in 2020 and 2021, while those 15 years of age or older reported an increase in the same period. Occupational intoxications with pesticides were within expected.

Keywords: SARS-CoV-2, forecasting, suicide attempt, drug addiction.

[I] Wednesday 17, 08:30-08:50, Room 3, session: LACSC session 3: Applications

[II] University of Costa Rica, Costa Rica, edgardo.gutierrez@ucr.ac.cr
[III] University of Costa Rica, Costa Rica, shuwei.chou@ucr.ac.cr
[IV] National Center for Poison Control, Children University, San José, Costa Rica, msomarrib@ccss.sa.cr

Data Science in Finance – Classification of Areas and Methods[I]

Communication

JAJUGA, KRZYSZTOF[II]

Poland

The paper gives an overview of the issues related to the application of statistical and data analysis methods in financial research. Firstly, there is classification of different areas in financial research where the quantitative methods are applied. It refers to such areas as for example: valuation of financial instruments and companies, analysis of prices and returns of financial instruments, term structure of interest rates, analysis of market risk, analysis of credit risk. Secondly, there is classification of data analysis methods, with the particular discussion on the interplay between classical statistical methods (including statistical learning) and machine learning methods (subset of artificial intelligence methods)

This leads to matching areas and methods. Finally, the challenges of the use of exploratory methods (both classical statistical learning and machine learning) are discussed, leading to model risk, which is understood as the application of erratic model in real world. To manage this type of risk, two tasks should be performed: assessment of reliability of data and assessment of adequacy of methods used to solve particular financial problem.

Keywords: data science, statistical learning, machine learning.

References

[1] Cao Larry (ed.) (2023), Handbook of Artificial Intelligence and Big Data Algorithms in Investments, CFA Institute Research Foundation. Dixon Matthew, Halperin Igor, Bilokon Paul (2020), Machine Learning in Finance. From Theory to Practice, Springer.

[2] Efron Bradley, and Hastie Trevor (2016), Computer Age Statistical Inference. Algorithms, Evidence and Data Science, Cambridge University Press, Cambridge.

[3] Jokhadze Valeriane, and Schmidt Wolfgang (2020), Measuring model risk in financial risk management and pricing. International Journal of Theoretical and Applied Finance 23: 2050012-1-37.

[4] Shalev-Shwartz Shai, and Ben-David Shai (2014), Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, Cambridge.

[I] Thursday 18, 10:40-11:00, Room 2, session: Data Science in Economics, Finance and Management 2

[II] Wroclaw University of Economics and Business, Poland, Krzysztof.Jajuga@ue.wroc.pl

Multivariate Clustering of Nonparametric Time Trends[I]

Communication

KHISMATULLINA, MARINA[II]

The Netherlands

In this paper, we propose a new clustering procedure for uncovering hidden group structure in multivariate time series based on the trends they exhibit. While clustering methodologies for univariate time series have been extensively studied, relatively few rigorous statistical methods have been developed for clustering multivariate time series data. However, the increasing volume of temporal data presents an opportunity to reveal meaningful hidden structures by clustering entities across multiple time series recorded simultaneously for each of the entities. The proposed clustering procedure is based on a new dissimilarity measure that stems from a recently developed multiscale testing method for comparison of univariate nonparametric trends (Khismatullina and Vogt, 2022). Building on this method, we extend the methodology to accommodate multivariate time series and propose a new distance measure that is able to capture the difference between the trends. We show that the clustering algorithm has the desired asymptotic properties and complement the project with a simulation study. For illustration purposes, we conduct an empirical study on the dataset that comprises information for 73 weather stations across Spain. We observe substantial differences when applying univariate versus multivariate clustering techniques, a finding that underscores the importance of adopting multivariate techniques to fully capture the complex dynamics inherent in multivariate time series data.

Keywords: multivariate time series , time series clustering, nonparametric statistics, model-based clustering.

References

[1] Khismatullina, M., Vogt, M. (2022). Multiscale comparison of nonparametric trend curves. arXiv preprint, doi: 10.48550/arXiv.2209.10841

[I] Tuesday 16, 17:20-17:40, Room 2, session: Time Series Analysis and Pattern Recognition

[II] Erasmus University Rotterdam, Erasmus School of Economics, Econometric Institute, The Netherlands, khismatullina@ese.eur.nl

Model-based Bi-clustering using Multivariate Poisson-Lognormal with General Block-diagonal Covariance Matrix and its Applications[I]

Poster Presentation

KRAL, CAITLIN[II]

Canada

Bi-clustering is a technique that simultaneously clusters observations and features (i.e., variables) in a dataset. This technique is used in bioinformatics to gain valuable insight. For example, biclustering gene expression data can help to simultaneously identify clusters of disease and non-diseased patients and the network of genes with distinct correlation patterns based on their expression values. While several Gaussian mixture models-based biclustering approaches currently exist in the literature for continuous data, approaches to handle discrete data have not been well researched. Extending biclustering approaches to discrete data is imperative as such data is commonly found within real world applications such as bioinformatics. Recently, multivariate Poisson-lognormal (MLPN) models have emerged as an efficient model for modelling multivariate count data. It arises from a hierarchical Poisson structure which allows for over-dispersion and correlation (both positive and negative). Here, we propose a MPLN model-based bi-clustering approach that utilizes a block-diagonal covariance structure to allow for a more flexible structure of the covariance matrix. We demonstrate the clustering performance of the proposed model for clustering both observations and features using simulated and real-world data.

Keywords: bi-clustering, multivariate Poisson-lognormal, bioinformatics.

[I]Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II]Carleton University, Canada, caitlinkral@cmail.carleton.ca

Green Bond Yield Determination with the use of Machine Learning Methods. Comparison with Conventional Bonds[I]

Communication

KUZIAK, KATARZYNA[II] Kaczmarczyk, Klaudia[III] Colak, Caner[IV]

Poland

The paper focuses on the similarities between green bonds and conventional bonds. A feature of green bonds is lower yields compared to conventional bonds with the same risk. The purpose of this study is to examine the determinants of the yields of selected bonds present in the global financial market. Among the financial characteristics, ESG rating was included as a determinant (Zhang et al. 2021; Immel et al. 2021). Macroeconomic factors, such as the consumer price index (CPI) or GDP growth rate, would also affect the size of the greenium (Cavallo and Valenzuela 2010; Ivashkovskaya and Mikhaylova 2020; Nanayakkara and Colombage 2019). There is no single set of determinants in the literature, as different studies have shown different results for different markets and countries, as well as types of markets. This study will use machine learning methods (e.g. based on gradient enhancement over decision trees - CatBoost). The findings will lead to a better understanding of the green bond market for investors, researchers, regulators and potential issuers.

Keywords: green bonds, yields, ESG rating.

References

[1] Cavallo, Eduardo, and Patricio Valenzuela (2010) The determinants of corporate risk in emerging markets: An option-adjusted spread analysis. International Journal of Finance & Economics **15**: 59–74.

[2] Grishunin Sergei, Bukreeva Alesya, Suloeva Svetlana, and Burova Ekaterina (2023) Analysis of Yields and Their Determinants in the European Corporate Green Bond Market, Risks **11**(1), 14

[3] Immel, Moritz, Britta Hachenberg, Florian Kiesel, and Dirk Schiereck (2021) Green bonds: Shades of green and brown. Journal of Asset Management **22**: 96–109

[4] Ivashkovskaya, Irina, and Anna Mikhaylova (2020) Do Investors Pay Yield Premiums on Green Bonds? Journal of Corporate Finance Research **14**: 7–21

[5] Nanayakkara, Madurika, and Sisira Colombage (2019) Do investors in green bond market pay a premium? Global evidence. Applied Economics **51**: 4425–37.

[I] Thursday 18, 10:20-10:40, Room 2, session: Data Science in Economics, Finance and Management 2

[II] Wroclaw University of Economics and Business Dept. of Financial Investment and Risk Management, Poland, Katarzyna.Kuziak@ue.wroc.pl

[III] Wroclaw University of Economics and Business Dept. of Financial Investment and Risk Management, Poland, klaudia.kaczmarczyk@ue.wroc.pl

[IV] Wroclaw University of Economics and Business, Master Student, Poland, 188612@student.ue.wroc.pl

[6] Zhang, Ran, Yanru Li, and Yingzhu Liu (2021) Green bond issuance and corporate cost of capital. Pacific-Basin Finance Journal **69**: 101626.

Fuzzy Clustering of Attributed Networks[I]

Communication

LABIOD, LAZHAR[II] Nadif, Mohamed[III]

France

The growing recognition of attributed networks as a crucial element in the field of data science stems from the growing abundance of this type of data, particularly within network contexts [1, 2, 3, 4]. Our contribution presents a novel approach to clustering attributed graphs. It proposes an objective function that uses regularized fuzzy clustering to enhance the quality of embeddings while ensuring effective clustering of nodes within graphs. Therefore, instead of treating feature information X and node topology W as distinct entities, we propose to base ourselves on an objective function that integrates embedding and fuzzy clustering. Using the characteristics of a low-rank subspace and fuzzy clustering, our method aims to capture the intricate connections between X andW, thus improving the robustness of clustering. Experiments are carried out on benchmark-attributed networks of different sizes to assess how well our algorithm performs compared to leading clustering methods designed for the same task.

Keywords: fuzzy clustering, embedding, attributed graphs.

References

[1] Fettal, C., Labiod, L., Nadif, M.: Simultaneous Linear Multi-view Attributed Graph Representation Learning and Clustering. In WSDM, pp. 303-311 (2023)

[2] Fettal, C., Labiod, L., Nadif, M.: Scalable Attributed-Graph Subspace Clustering. In AAAI, pp. 7559-7567 (2023)

[3] Labiod, L.,Nadif, M.: PowerAttributed Graph Embedding and Clustering. IEEE Trans.Neural Networks Learn. Syst. 35(1): 1439-1444 (2024)

[4] Riverain, P., Fossier, S., Nadif, M.: Model-based Poisson co-clustering for Attributed Networks. ICDM (Workshops), pp. 703-710 (2021)

[I]Wednesday 17, 10:40-11:00, Room 1, session: Clustering, Classification and Discrimination 6 (A. Grané)

[II]Centre Borelli, UMR9010, Université Paris Cit, France, lazhar.labiod@u-paris.fr

[III]Centre Borelli, UMR9010, UniversitÃ©, Paris Cite, France, mohamed.nadif@u-paris.fr

Multiblock Regularized Least-squares Latent Variable Method[I]

Communication

LE, THU TRA[II]

The Netherlands

The next-generation approach to behavioral research relies on intensive data collection from multiple disciplinary domains. Behavior and cognition are no longer studied from the psychological perspective only but also from other disciplinary perspectives such as environmental, social, clinical, and biomolecular. This often leads to so-called high-dimensional multiview data. In analyzing this type of data, it is of great importance to disentangle distinct mechanisms underlying each data block from common mechanisms shared by all (or multiple) data blocks. Current latent variable methods are not appropriate to address this challenge. To this end, we propose a Multiblock Regularized Least-squares Latent Variable Method. The method uses hard cardinality constraint (instead of a penalized approach such as the group lasso) to impose sparsity across and within data blocks. That is, the model is estimated under the constraint that exactly C blocks of loadings are equal to zero to identify specific and shared mechanisms. In addition, within each data block, exactly K loadings are imposed to be zero to encourage variable selection to ease interpretation. Both latent variable scores and loadings are estimated in an alternating optimization scheme. The performance of the proposed method is evaluated in an extensive simulation study. We also demonstrate the use of the method using a real-world dataset.

Keywords: cardinality constraint , dimension reduction, multiblock data, latent variable, structural equation modeling.

[I]Monday 15, 17:25-17:25, Room 1, session: Dimension Reduction

[II]Tilburg University (Department of Methodology and Statistics), The Netherlands, t.t.le_1@tilburguniversity.edu

On Relation Between Separable Effects, Natural Effects, and Interventional Effects[I]

Poster Session

LIN, SHENG-HSUAN[II]

Japan

This paper provides a thorough comparison of causal models used in causal mediation analysis for identifying mediation effects, focusing particularly on separable effect method. Separable effects, proposed by James Robins, are claimed identifiable without the need for the untestable cross-world assumption implied by the nonparametric structural model with independent errors. However, our study clarifies that separable effects do not guarantee a mediation interpretation due to violation of a mediation null criteria. This paper reveals that when the separable effects provide mediation interpretation, they are the same as the natural effects, but rely on assumptions stronger than the cross-world assumption. Furthermore, when there is a mediator-outcome confounder affected by exposure, separable effects do not have a mediation interpretation, while another study has shown that the natural effects and interventional effects have a mediation interpretation under a no conditional mean causal interaction assumption. Our study elucidates the relationship between separable, natural, and interventional effects and proposes an integrated framework that is applicable to practical analysis of clinical studies. We emphasize the key role of the untestable isolation assumptions in separable effects for mediation interpretation, and highlight a trade-off between the interpretability and falsifiability of assumptions.

Keywords: causal mediation analysis, separable effects, natural effects, interventional effects, recanting witness, casual model.

[I]Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II]Institute of Statistics, National Yang Ming Chiao Tung University, Japan,

Understanding Omics Links Behind Glioma Heterogeneity: A Network and Clustering Approach[I]

Communication

LOPES, MARTA B[II]

Portugal

Gliomas are primary malignant brain tumors known for their generally poor prognoses, largely due to the molecular heterogeneity observed across different tumor types. In this study, we propose a comprehensive strategy involving network discovery and clustering to explore the transcriptomics landscape of gliomas and support targeted therapeutic research on the potential biomarkers identified. In a first stage of the methodology proposed, the graphical lasso [1] algorithm is applied to disclose interactions among genes in each glioma type through the estimation of sparse transcriptomics networks. Centrality measures and modularity detection [2] are then used to aid in the identification of key genes in glioma types. In the second stage, spectral clustering [3] of patient similarity networks is applied to evaluate the suitability of the genes identified in grouping patients into the glioma types. The results obtained underscore the potential of the proposed approach in uncovering relevant genes associated with glioma heterogeneity. Further research efforts may involve the biological validation of the disclosed network insights on glioma heterogeneity.

Keywords: glioma, omics, network, clustering.

References

[1] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441 (2008)

[2] John, C.R., Watson, D., Barnes, M.R., Pitzalis, C., Lewis, M.J.: Spectrum: fast density-aware spectral clustering for single and multi-omic data. Bioinformatics 36(4), 1159–1166 (2020)

[3] Newman, M.E.J.: Modularity and community structure in networks. PNAS 103(23), 8577–8582 (2006)

[I] Monday 15, 17:45-18:05, Room 3, session: Big Data and High-Dimensional

[II] Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, Portugal, marta.lopes@fct.unl.pt

A Collection of Biplots for Classification[I]

Plenary Talk

LUBBE, SUGNET[II]

South Africa

In a multidimensional world it is important to visualise our data in two or three dimensions. Especially in a classification setting a picture is worth a thousand words in illustrating how (well) classes are separated and how they overlap. Biplots, as the prefix "bi-" suggests, represent the cases, grouped into classes, as well as the variables. By having a visual display of the multiple variables, interpretation on which variables separate which classes to what extent can be easily obtained.

After a brief general introduction on biplot methodology we will step through an array of options related to classification settings. In traditional multivariate statistics courses, classification is introduced through linear discriminant analysis. We will start from this point onward and look at biplot options for linear and nonlinear classification, continuous and categorical data, and make our way to more recently developed classification methodology.

Keywords: biplot , classification, linear discriminant analysis.

[I]Thursday 18, 17:00-18:00, Auditorium, session: Closing Plenary Talk

[II], South Africa,

Predicting Soil Bacterial and Fungal Communities at Different Taxonomic Levels Using Machine Learning[I]

Communication

MAKARENKOV, VLADIMIR[II] Aouabed, Zahia[III] Bouaoune, Mohamed Achraf[IV]
Hijri, Mohamed[V]

Canada

It is widely known that predictions about macrobiological communities depend on the taxonomic scale. Nevertheless, the applicability of such predictions remains uncertain when extended to microbial communities of the soil. This study employs various traditional machine learning techniques to forecast bacterial and fungal communities within the soil across different taxonomic levels. To investigate this avenue, we use an extensive soil microbiome dataset collected by diverse research groups. Our bacterial results indicate significantly superior prediction accuracy at the Phylum, Class, and Order taxonomic levels compared to the Family and Genus levels. Lower prediction scores, compared to bacteria, were generally found for fungi, with the best results obtained at the Phylum and Class taxonomic levels. Overall, our findings suggest a consistent trend across taxonomic scales, bridging macrobiological and soil microbiological communities. For bacterial data, our prediction results obtained using the Random Forest and Gradient Boosting methods were generally better than those found by Averill and co-authors, who used the Dirichlet multivariate regression model in their study recently published in Nature Ecology and Evolution. For fungal data, we recommend using Random Forest to provide the soil community predictions.

Keywords: biological data prediction , linear regression, decision trees, random forest, gradient boosting.

[I] Thursday 18, 10:00-10:20, Room 1, session: Machine Learning

[II] University of Quebec in Montreal, Canada, makarenkov.vladimir@uqam.ca
[III] Computer Science, Université du Québec à Montréal, Canada, aouabed.zahia@uqam.ca
[IV] Computer Science, Université du Québec à Montréal, Canada, mohamed.achraf.bouaoune@umontreal.ca
[V] Biological Sciences, Université du Québec à Montréal, Canada, mohamed.hijri@umontreal.ca

Robust Estimation of the Range-based GARCH Model: Forecasting Volatility, Value at Risk and Rxpected Shortfall of Cryptocurrencies[I]

Communication

MALECKA, MARTA[II] Fiszeder, Piotr[III]

Poland

We combine the range-based GARCH model with the modified robust method of estimation and suggest a new approach to model volatility of returns. Thanks to this merger, we use more information which are commonly available alongside daily closing prices, i.e., low and high prices but at the same time we limit the influence of extreme observations on the estimation results. Owing to this, the procedure is not as sensitive to outliers as the maximum likelihood estimation of the range-based models. We also propose to introduce the change to the robust method, which adds elasticity in treating the outliers and serves to reflect the observations of financial markets, where, after occurrence of outliers, the volatility persists at an increased level. We apply this method to five selected cryptocurrencies: Bitcoin, Ethereum Classic, Ethereum, Litecoin and Ripple. The forecasts of variance based on the proposed approach are more accurate than forecasts from three benchmarks: the standard GARCH model, the standard range-based GARCH model and the GARCH model with the robust estimation.

Keywords: range-based GARCH model, robust estimation methods, cryptocurrency volatility.

[I]Thursday 18, 08:50-09:10, Room 2, session: Data Science in Economics, Finance and Management

[II]Faculty of Economics and Sociology, University of Lodz, Lodz, Poland, Poland, marta.malecka@uni.lodz.pl

[III]aculty of Economic Sciences and Management, Nicolaus Copernicus University in Torun, Torun, Poland, piotr.fiszeder@umk.pl

Clustering for High-Dimensional, Nested Data with Categorical Outcomes Using a Generalized Linear Mixed Effects Model with Simultaneous Variable Selection[I]

Poster Presentation

MANNING, SAMANTHA[II]

United States

I propose a model-based clustering method for high-dimensional, longitudinal data with categorical outcomes via regularization. The development of this method was motivated in part by a study on 177 Thai mother-child dyads to identify risk factors for early childhood caries (ECC). Another considerable motivation was a dental visit study of 308 pregnant women to ascertain determinants of successful dental appointment attendance. There is no available method capable of clustering longitudinal categorical outcomes while also selecting relevant variables. Within each cluster, a generalized linear mixed-effects model is fit with a convex penalty function imposed on the fixed effect parameters. Through the expectation-maximization algorithm, model coefficients are estimated using the Laplace approximation within the coordinate descent algorithm, and the estimated values are then used to cluster subjects via k-means clustering for longitudinal data. The Bayesian information criterion can be used to determine the optimal number of clusters and the tuning parameters through a grid search. My simulation studies demonstrate that this method has satisfactory performance and is able to accommodate high-dimensional, multi-level effects as well as identify longitudinal patterns in categorical outcomes.

Keywords: model-based clustering, modelling high-dimensional and complex data, generalized linear models, mixed-effects models.

[I]Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II]Department of Biostatistics and Computational Biology at the University of Rochester Medical Center, United States, samantha_manning@urmc.rochester.edu

Traffic Accidents in Nicaragua[I]

Communication

MARADIAGA RIVAS, ERICKA MARÍA[II] Rostrán Molina, Ana Cristina[III]

Soto Pineda, Ángel[IV]

Nicaragua

The World Health Organization (2022), traffic accidents are a public health issue. The United Nations (2023) points out that in developing countries, 90% of traffic accidents are the leading cause of death among children and young people aged 5 to 29. These accidents result in annual losses of 2% to 5% of Gross Domestic Product and cause the deaths of 1.3 million and injuries to 50 million people. The total number of traffic accidents that occurred in Nicaragua from 2010 to 2021 was analyzed, as reported by the Nicaraguan National Police and recorded in the database of the Institute of Development Information (INIDE). During the study period, traffic accidents increased by 93%. The trend of the total number of accidents and total number of injuries is decreasing. However, the total number of fatalities is increasing; therefore, when accidents occur, they tend to be fatal. The multiple regression model with the total number of accidents as the endogenous variable is explained by the age of those involved in traffic accidents and the total number of injuries by gender. This is significant (P=0.000; a=0.05) The variation in one percent of the total number of accidents will be an increase of 96% in traffic accidents, the variation in one percent of the age of traffic accidents will increase by 5.7% (ceteris paribus) The regression model with a total dependent variable of injuries is significant (P=0.000; a=0.05). Age is a variable sensitive to the change of signs. Male drivers, passengers, and pedestrians are the largest injured victims.

Keywords: accidents, traffic, model.

References

[1] Asamblea Nacional de la República de Nicaragua. (27 de mayo de 2014). Ley No.. 431 "Ley para el régimen de circulación vehicular e infracciones de tránsito", con sus reformas incorporadas. La Gaceta, págs. 1–41.

[2] Box G, & Jenkins, G. (1970). Time series analysis: forecasting and control. San Francisco: Holden Day. Escuela Politécnica Nacional. (2021). Modelización Econométrica de los Accidentes de Tránsito en el Ecuador. Revista Politécnica, pag. 1.

[3] Herrera Briones, J., Mira McWilliams, J., & Sanjuro de No, M. A. (2021). Anális y predicción de la lesividad en accidentes de tráfico mediante la aplicación de random forest. Madrid: Universidad Politécnica de Madrid Escuela Técnica Superior de Ingenieros Industriales.

[I] Wednesday 17, 09:10-09:30, Room 3, session: LACSC session 3: Applications

[II] Universidad Nacional Autónoma de Nicaragua (UNAN–León), Nicaragua, ericka.maradiaga120@est.unanleon.edu.ni

[III] Universidad Nacional Autónoma de Nicaragua (UNAN–León), Nicaragua, anarostran@ct.unanleon.edu.ni.

[IV] Universidad Nacional Autónoma de Nicaragua (UNAN–León), Nicaragua, Daniel.angel.soto120@est.unanleon.edu.ni

[4] INIDE. (2022). Anuario Estadistico 2022. Managua,Nicaragua: INIDE.

[5] Mayoral Grajeda , E. F., Cueva Colunga, A. C., Pérez Castro, J. G., & Mendoza Díaz , A. (2015). Análisis de la siniestralidad de los usuarios vulnerables en carreteras federales. San Fandila: Instituto Mexicano de Transporte Publicación Técnica No. 453.

[6] OMS. (23 de 08 de 2018). Stio Web mundial de la Organizacion Nacional de la Salud. Obtenido de Genero y Salud: https://www.who.int

[7] OMS. (2021). Plan mundial para el decenio de acción para la seguridad vial 2021–2030. Ginebra: OMS.

[8] OMS. (06 de 04 de 2022). Centro de Prensa. Obtenido de Traumatismos causados por el tránsito: https://www.who.int/es

[9] ONU. (02 de 06 de 2023). El cinturón de seguridad ha salvado millones de vidas en los últimos 50 años. SEGURIDAD VIAL.

[10] OPS. (07 de nov de 2016). La seguridad vial en la región de las Américas. Washington, D.C.: OPS OMS. Obtenido de La seguridad vial en la Región de las Américas.

[11] OPS. (2023). Implementación de medidas de seguridad vial prioritarias en América Latina y el Caribe. Washington: OPS.

[12] PLN. (02 de septiembre de 2022). Policía Nacional . Obtenido de Plan de emergencia vial. Managua: https://www.policia.gob.ni/

[13] Policía Nacional de Nicaragua. (02 de septiembre de 2022). Plan Nacional de Emergencia Vial. Managua: Policia Nacional. Obtenido de Policía Nacional de Nicaragua: https://www.policia.gob.ni/wp–content/uploads/2022/09/Plan–Nacional–de–Emergencia–Vial–Aprobada.pdf

[14] Senisse, A. (2016). La seguridad vial en la región de las Américas. Washington, D.C.: ISBN: 978–92–75–11912–9.

Unsupervised Methods for the Creation of Orthonormal Bases in Compositional Data: R-mode Clustering[I]

Communication

MARTÍN FERNÁNDEZ, JOSÉ ANTONIO[II]

Spain

R-mode hierarchical clustering (HC) identifies interrelationships between variables which are useful for variable selection and dimension reduction. The application of HC in R-mode to Compositional Data (CoDa) must be consistent with the fundamental properties of the compositional geometry, also known as the Aitchison geometry. A composition is a multivariate quantitative description of the parts or components of a whole conveying relative information, commonly expressed as a vector of proportions. The critical element of the Aitchison geometry is the inner product defined via the log-ratio coordinates of the compositions. This geometry allows to express a composition as coordinates in an orthonormal basis, formed by log-ratios and called olr-coordinates. Recent publications introduce R-mode agglomerative HC methods in CoDa for creating orthonormal log-ratio basis. The HC methods form hierarchical groups of mutually exclusive subsets of parts which can be associated to a sequential binary partition of the parts. In this talk, we explore the basic concepts of the R-mode clustering algorithms and the connections between concepts such as distance between parts, cluster representative of a group of parts, and compositional biplot. Practical examples will be presented to visually illustrate the proposed approach.

Keywords: compositional data , logratio, simplex, R-mode clustering.

[I]Wednesday 17, 08:50-09:10, Room 1, session: Clustering, Classification and Discrimination 5 (A. Grané)

[II]Universitat de Girona, Spain, josepantoni.martin@udg.edu

Spatial Agent-Based Model for Aedes Aegypti Mosquitoes in the Urban Area in Arica (Chile)[I]

Poster presentation

MARTÍNEZ, DIANA MARCELA[II] Martínez, Kerlyns[III] Velandia, Daira[IV]

Chile

The spread of the Aedes *aegypti* mosquito, which transmits the dengue, is an important public health issue. One approach to estimating the parameters of dengue epidemics is the agent-based spatial model, such as the MOMA model developed by Mannerat and Daude in 2006, which simulates the interaction between mosquitoes and their environment. It is limited to the study of mosquito dynamics and does not assess dengue dynamics. To this end, the aim of this study was to implement an agent-based model of Aedes *aegypty* for the estimation of dengue epidemic parameters under specific urban social, geographic, and climatic conditions (in Arica- Chile). The model includes three agent types: a female Aedes agent, a spatial object agent representing the physical environment, and world agent representing the simulation environment. The Aedes agent includes entomological and behavioral parameters, while the spatial object is constructed using data on population density, land use, and satellite imagery. During the simulation, there is a fixed time step. Within this time step, Aedes agents can observe their environment and choose targets that satisfy their needs, such as biting, taking nectar, laying eggs, and resting. They prioritize their needs using a behavioral decision-making process. We simulated a dynamic population based on social classification and found a strong relationship between mosquito density and flight patterns, urban topology, and human behavior and density.

Keywords: agent-based model, mosquito population dynamic, spatial simulation.

References

[1] Maneerat, S., & Daude, E. (2016). A spatial agent-based simulation model of the dengue vector Aedes aegypti to explore its population dynamics in urban areas. Ecological Modelling, 333, 66–78.

[I] Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II] Universidad de Valparaíso, Instituto de Estadística, Chile, diana.martinez@postgrado.uv.cl

[III] Universidad de Valparaíso, Instituto de Estadística, Chile, kerlyns.martinez@uv.cl

[IV] Universidad de Valparaíso, Instituto de Estadística, Chile, daira.velandia@uv.cl

Nicaragua Migrations: Origin–Destiniy[I]

Communication

MARTÍNEZ, LARISSA[II] Rostrán Molina, Ana Cristina[III]
Betanco Corea, Gloria Stephany[IV]

Nicaragua

The International Organization for Migration (IOM, 2021), states that an estimated 281 million people were living in a country other than their home country in 2020. They also state that international migrants represented 3.6% of the world's population. Migration trends were analyzed in Nicaragua origin destination, with the databases of the: National Institute of Development Information (INIDE), Central Bank of Nicaragua (BCN), IOM and World Bank (WB) from 2006 to 2020. The maximum number of Nicaraguan migrants occurred in 2017 with 988,413 people; from that year the trend of migration, decreases, by 2020 there were 325,092 people. The main destination countries of Nicaraguan migration are: Costa Rica with 43%-75%, United States 12%–20% and Honduras with 0.3% to 33%. The model was estimated to explain the migration of Nicaraguans as a function of remittances received and the GDP per capita of the U.S.A. The model is significant (p=0.000). Nicaraguan migration is incentivized by U.S. GDP per capita. Remittances received in Nicaragua generate a negative contribution to migration under the (ceteris paribus) assumption.

Keywords: migration, regression, origin, destiny.

References

[1] BCN. (26 de Noviembre de 2022). Base de Datos Estadisticos. Obtenido de Banco Central de Nicaragua: https://www.bcn.gob.ni/base-de-datos-estadisticos

[2] BM. (Septiembre de 2023). Datos. Obtenido de Banco Mundial: https://datos.bancomundial.org/

[3] BM. (2023). Migrantes, refugiados y sociedades. BM. Washington, DC: Banco Internacional de Reconstrucción y Fomento/Banco Mundial. Obtenido de https://www.bancomundial.org/es/home

[4] CEPAL. (Enero de 2019). Observatorio Demográfico de AmÃ©rica Latina y el Caribe. Obtenido de Migración Internacional.: https://repositorio.cepal.org

[5] CEPAL. (2021). Introducción a la desigualdad. Comisión Económica para AmÃ©rica Latina y el Caribe. CEPAL,ONU. Obtenido de https://igualdad.cepal.org/es

[6] CEPAL,CELADE. (2006). Migración internacional Y desarrolllo en Nicaragua. Santiago de Chile: CELADE.

[I] Wednesday 17, 08:50-09:10, Room 3, session: LACSC session 3: Applications

[II] Universidad Nacional Autónoma de Nicaragua (UNAN-León) Nicaragua, larissa.martinez120@est.unanleon.edu.ni,
[III] Universidad Nacional Autónoma de Nicaragua (UNAN-León) Nicaragua,anacrostran@ct.unanleon.edu.ni.
[IV] Universidad Nacional Autónoma de Nicaragua (UNAN-León) Nicaragua, III. gloria.betanco120@est.unanleon.edu.ni

[7] INIDE. (Septiembre de 2022). Anuarios Estadisticos. Obtenido de Instituto Nacional de Información de Desarrollo: https://www.inide.gob.ni/Home/Anuarios

[8] Islas Camargo, A., & Moreno Santoyo, S. (2011). Determinantes del flujo de remesas en MÃ©xico, un análisis empírico. (EconoQuantum, Ed.) 7(2), 9-36.

[9] OIM. (Junio de 2017). Organización Internacional para las Migraciones. Obtenido de https://www.iom.int

[10] OIM. (2018). Informe sobre las Migraciones en el mundo 2019. Ginebra: OIM.

[11] OIM. (2019). Informe sobre las Migraciones en el Mundo 2020. Organización Internacional para las Migraciones. Ginebra: OIM. Obtenido de Organización Internacional para las Migraciones. OIM.

[12] OIM. (2020). Informe sobre las Migraciones en el mundo 2019. Ginebra: OIM.

[13] OIM. (2021). Indicadores de Gobernanza de la Migración República de Nicaragua. . Ginebra: OIM. Obtenido de https://publications.iom.int/books/indicadores-de-gobernanza-de-la-migracion-perfil-2021-republica-de-nicaragua

[14] OIM. (2021). Informe sobre las Migraciones en el Mundo 2022. En OIM (Ed.). Ginebra.

[15] OIM. (2021). Organización Internacional para las Migraciones. Obtenido de Por una migración benÃ©fica para todos: https://www.iom.int

[16] OIM. (2023). Missing migrants project. Obtenido de https://missingmigrants.iom.int/

[17] Torres Betanco, N., & Aráuz Torres , M. (2023). Remesas en Nicaragua y su influencia en el mercado laboral. Revista de Economia y Finanzas, 10, P4.

[18] U.S. Border Proteccion. (2023). An official website of the States govermnent. Obtenido de http://www.cpb.gov

[19] UIP, OIT & ONU. (2015). Migración, Derechos Humanos y Gobernanza: Manual para Parlamentarios NÂ° 24. Unión Interpalamentaria, Ginebra.

Multiblock Methods for Learning Structural Equation Models: An Overview[I]

Communication

MARTÍNEZ-RUIZ, ALBA[II]

Chile

Structural equation models (SEM) are powerful tools for estimating the value of a set of unobserved or latent variables. These latent variables represent concepts that can be inferred from blocks of variables measured in the same collection of individuals. The nature of the association between latent and manifest variables "formative or reflective, composites or factors" determines the mathematical model for estimating the variables and the relationship between them. Several methods exist for learning SEM, with sequential multiblock component methods widely applied in practice. A general framework is provided by Regularized Generalized Canonical Correlation Analysis (RGCCA). The machinery involves the construction of a set of block components in such a way to maximize a function of the covariances between linear combinations of the block of variables. Special cases of RGCCA are the classic methods of canonical correlation analysis and redundancy analysis, as well as SUMCOR, SSQCOR, and SABSCOR. In this work, I review the main definitions related to RGCCA, including its origins, theoretical foundations, optimization problems, and algorithms.

Keywords: structural equation models , RGCCA, multiblock learning.

[I]Tuesday 16, 08:30-08:50, Room 3, session: LACSC: Data Science and Computational Statistics: Theory and applications

[II]Universidad Diego Portales, Chile, alba.martinez.ruiz@gmail.com

Alternating Least Squares Algorithm: Speedup versus Accuracy[I]

Communication

MARTÍNEZ-RUIZ, ALBA[II] Niang, Ndèye[III] Lemus Henríquez, Pablo[IV]

Chile

We propose a modification of the classic alternating least squares (ALS) algorithm to reduce the dimensionality of third-order tensors. The ALS algorithm enables the simultaneous calculation of components matrices for each dimension by minimizing a loss function. The algorithm involves the factorization of the matricized tensor in the iterative process. The suggested extension adapts the algorithm by performing the decomposition of partitioned unfolding matrices. The adaptation speeds up the analysis of third-order high-dimensional tensors and enables the estimation of the component matrices for each dimension much more efficiently, although with a certain cost in accuracy. We demonstrate the properties of the new algorithm by performing numerical experiments on synthetic data and a real-data application.

Keywords: tensor decomposition, alternating least squares (ALS), dimensionality reduction.

[I] Tuesday 16, 08:30-08:50, Room 3, session: LACSC: Data Science and Computational Statistics: Theory and applications

[II] Diego Portales University, Chile, alba.martinez.ruiz@gmail.com

[III] Cédric-CNAM, France, keita@cnam.fr

[IV] Diego Portales Univesity, Chile, pablo.lemus@mail.udp.cl

Statistical Science Meets Digital Health: Distributional Data Analysis in Digital Health[I]

Tutorial

MATABUENA, MARCOS[II]

United States

Digital health is transforming clinical practice and driving the precision medicine paradigm to personalize healthcare. In this context, the new information collected is of a functional nature over time, and patients are monitored in real-life conditions. Therefore, classical time series techniques cannot be used for many modeling tasks, and users must focus on comparing the distributional differences of the biological time series. Distributional data analysis is a powerful tool for analyzing and validating new biomarkers emerging in digital health data.

The objective of this tutorial is to introduce the main tools of distributional data analysis for analyzing clinical data in digital health. The session consists of two parts. The first part is theoretical, where the main data analysis tools are introduced. The second part, using R code, demonstrates with real cases how to use these techniques in practice with relevant examples of wearable data.

Keywords: digital health, distributional data analysis, precision medicine.

[I]Monday 15, 08:30-12:00, Room 2, session: Tutorial 2

[II]Postdoctoral researcher in the department of biostatistics at Harvard University., United States,

Statistical Science Meets Digital Health[I]

Plenary Talk

MATABUENA, MARCOS[II]

United States

Jen-Hsun Huang, CEO of NVIDIA, recently stated that "digital biology will be one of the greatest revolutions in human history." In this era of data-driven systems, statistical methods must evolve to meet the analytical demands emerging from various tasks in digital health. These advancements are essential to increase the efficiency of healthcare systems worldwide and address current public health problems such as diabetes and the aging of modern societies. One key area in the transition to digital health solutions in clinical practice is the discovery and validation of digital biomarkers, which often exist as latent random variables in metric spaces.

The main objective of this talk is to demonstrate how digital health can inspire the creation of novel statistical methods, such as uncertainty quantification techniques in metric spaces, and how the statistical modeling of these complex objects can lead to improved decision-making in modern clinical practice. These novel biomarkers can include graphs and probability distributions of biological times, for example, from continuous glucose monitoring.

Keywords: digital health, biostatistics, precision public health, conformal prediction.

[I]Tuesday 16, 13:30-14:30, Auditorium, session: Plenary Talk 3

[II]Postdoctoral researcher in the department of biostatistics at Harvard University., United States,

Spatio-temporal Hierarchical Clustering of Interval Time Series with Application to Suicide Rates in Europe[I]

Communication

MATTERA, RAFFAELE[II] Franses, Philip Hans[III]

Italy

In this paper we investigate similarities of suicide rates in Europe, which are available as interval time series. For this aim, a novel spatio-temporal hierarchical clustering algorithm for interval time series data is proposed. The spatial dimension is included in the clustering process to account for possible relevant information such as weather conditions, sunlight hours and socio-cultural factors. Our results indicate the presence of six main clusters in Europe, which almost overlap with the sunlight hours distribution. Differences between male and female suicide rates are also investigated.

Keywords: symbolic data analysis , spatio-temporal modelling, spatial data science.

References

[1] Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2018). ClustGeo: an R package for hierarchical clustering with spatial constraints. Computational Statistics, 33(4), 1799-1822.

[2] Maharaj, E. A., Teles, P., & Brito, P. (2019). Clustering of interval time series. Statistics and Computing, 29, 1011-1034.

[3] Mattera, R., & Franses, P. H. (2023). Are African business cycles synchronized? Evidence from spatio-temporal modeling. Economic Modelling, 128, 106485.

[I] Tuesday 16, 10:00-10:20, Room 2, session: Symbolic Data Analysis 2

[II] Department of Social and Economic Sciences, Sapienza University of Rome, Italy, Italy, raffaele.mattera@uniroma1.it

[III] Econometric Institute, Erasmus University of Rotterdam, the Netherlands, The Netherlands, franses@ese.eur.nl

Improving Functional Classification Performance through Diversity: The Functional Voting Approach[I]

Communication

MATURO, FABRIZIO[II] Riccio, Donato[III] Romano, Elvira[IV]

Italy

In recent decades, Functional Data Analysis (FDA) has become widely popular as a framework for analyzing data that are inherently functions in the domain of time. Although supervised classification has been extensively explored in recent decades within the FDA literature, ensemble learning of functional classifiers has only recently emerged as a topic of significant interest. The focal point of this study lies in the realm of ensemble learning for functional data and aims to show how different functional data representations can be used to train ensemble members and how base model predictions can be combined through majority voting. The so-called Functional Voting Classifier (FVC) is proposed to demonstrate how diversity can increase predictive accuracy. The framework presented provides a foundation for voting ensembles with functional data and can stimulate a highly encouraging line of research in the FDA context.

Keywords: fda, classification, voting, diversity.

[I] Thursday 18, 10:00-10:20, Room 3, session: Functional Data Analysis 2

[II] Universitas Mercatorum, Faculty of Technological and Innovation Sciences, Rome, Italy, fabrizio.maturo@unimercatorum.it

[III] University of Campania Luigi Vanvitelli, Machine Learning Engineer and Student in the Data Science Master's Degree Program, Caserta, Italy, donato.riccio@studenti.unicampania.it

[IV] University of Campania Luigi Vanvitelli, Department of Mathematics and Physics, Caserta, Italy, elvira.romano@unicampania.it

Optimization Strategies for Bioprocess Parameterization: A Comparative Evaluation[I]

Communication

MEDL, MATTHIAS[II]

Austria

Operating biopharmaceutical manufacturing processes at optimal con-ditions is critical to maximize productivity and product quality, while minimizingmanufacturing cost and environmental impact. However, the experimental budgetfor the search of optimal bioprocess parameters is limited. Thus, it is essential toemploy efficient and effective optimization strategies such as Design of Experimentsor Gaussian Process optmization. In practice, these optimization strategies have tobe parameterized. It is unclear under which circumstances what optimzation strat-gey performs best and how to paramterize it effectively. The aim of this study is tocompare the performance and robustness of multiple optimization strategies acrossa diverse range of parameter and process configurations of a simulated ion exchangechromatography process. The outcome of this study is to provide guidance navigatingthe large decision space encountered when performing (bio)process optimization.

Keywords: parameter optimization, design of experiments, gaussian process optimization, biorprocess.

[I] Tuesday 16, 14:50-15:10, Room 3, session: Applications

[II] University of Natural Resources and Life Sciences Vienna, Austria, matthias.medl@boku.ac.at

A Quantile Extension to Functional PCA[I]

Communication

MÉNDEZ CIVIETA, ÁLVARO[II]

Spain

This study presents the Functional Quantile Principal Component Analysis (FQPCA). This methodology draws on the probabilistic approach for PCA proposed by [1] and extends the functional PCA to the quantile regression framework [2]. This results in a model that describes the full curve-and time-specific probability distribution that underlies individual measurements, estimating smooth, curve-specific quantile functions that are dependent on a set of principal components. The median can be seen as a robust alternative of the mean provided by traditional FPCA, while other quantiles give a more complete understanding of the subject- and time- specific data distribution, and may be particularly useful when distributions are skewed, heteroscedasti or vary across subjects. The necessity for this methodology is demonstrated by our illustrative example: we examine the physical activity level of over 3600 individuals in a single day using accelerometer data from the National Health and Nutrition Examination Survey (NHANES) and are able to compare information from different quantile levels. The proposed methodology is available as a package in R programming language.

Keywords: accelerometer data , functional data, quantile regression, PCA.

References

[1] C. M. Bishop. Bayesian PCA. In: Advances in Neural Information Processing Systems 11 (1999), 382–388.

[2] R. Koenker; & G. Bassett, Jr. Regression Quantiles. Econometrica, Vol. 46, No. 1 (Jan., 1978), 33–50.

[I] Thursday 18, 08:50-09:10, Room 3, session: Functional Data Analysis

[II] Department of biostatistics, Columbia University, New York, Spain, am5490@cumc.columbia.edu

Bridge the Gap Between Gradual Patterns and Statistical Correlations[I]

Communication

MEPHU NGUIFO, ENGELBERT[II] Maureen Domche, Norbert Tsopze[III] Jerry Lonlac[IV]

France

Gradual patterns mining aims to extract from numerical data, the frequent covariations between attributes (or variables) x_i of the form "The more/less $x_1, ...,$ the more/less x_1". Gradual patterns are significant for capturing the variability of numerical values in applications when the volume of data becomes large. Moreover, statistical correlation, which highlights relationships between variables, can also be used to express covariations in numerical data.

Although gradual patterns and statistical correlations capture co-variations from numerical data, gradual patterns provide more expressive knowledge. To our knowl- edge, no work in the literature focused on studying the differences between these two concepts to highlight the limits and the advantages of gradual patterns regarding statistical correlations.

In this work, we conduct a comparative study between the gradual patterns extracted using different semantics of graduality and statistical correlations, presenting the similarities, differences, advantages and disadvantages of each of the concepts for numerical data processing.

This study is completed by experiments carried out on several numerical databases, the results of which confirm the contribution of gradual patterns compared to statistical correlations.

Keywords: pattern mining, gradual patterns, correlation analysis.

References

[1] Domche, M., Lonlac, L., Tsopz, N., Mephu Nguifo, E. (2024).Une étude comparative entre Motifs Graduels et Corrélations Statistiques. In EGC 2024, vol. RNTI-E-40, pp.333-334

[I] Tuesday 16, 17:00-17:20, Room 1, session: Data Mining

[II] University Clermont– Auvergne, Clermont Auvergne INP, CNRS, France, engelbert.mephu_nguifo@uca.fr

[III] Departement d'Informatique, Université de Yaound'e 1, Cameroun, e-mail: {maureennouno2000, tsopze.norbert}@gmail.com

[IV] IMT Nord Europe, IMT, Université de Lille, CERI SN, Lille, 59653, France, e-mail: jerry.lonlac@imt-nord-europe.fr

(Data oblivious) Random Projections for (data aware) Model-based Clustering[I]

Plenary Talk – Presidential Address

MONTANARI, ANGELA[II]

Italy

Classical model-based clustering methods show a disappointing behaviour in high dimensional space. In this case, the over-parametrization issue is typically solved by neglecting or excessively simplifying the correlation structure of the data.

Random projections (RPs) have shown to provide promising results in the context of high-dimensional multivariate analysis. In this talk, the problem of clustering high dimensional data is addressed by resorting to a RP-based ensemble of low-rank estimates for the group-specific covariance matrix.

The performances of the proposal are assessed in terms of both clustering accuracy and estimate precision through numerical studies and real data applications. This is a joint work with Laura Anderlucci and Silvia Dallari.

Keywords: high-dimensional clustering, random projections, covariance matrix estimation.

References

[1] L. Anderlucci, F. Fortunato, A. Montanari: High-dimensional clustering via Random Projections. Journal of Classification. **39**, 191–216 (2022).

[2] Marzetta, T. L., Tucci, G. H., Simon, S. H.: A random matrix-theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory. **57**(9), 6256–6271 (2011).

[3] Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E.: Model-based clustering, classification, and density estimation using mclust in R. Chapman and Hall/CRC (2023)

[I] Wednesday 17, 11:10-12:10, Auditorium, session: Plenary Talk 5

[II] Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, angela.montanari@unibo.it

Using Polynomials to Explain Classification Outputs from Neural Networks[I]

Communication

MORALA, PABLO[II]

Spain

Neural networks, specially with the advent of deep learning, have shown and outstanding performance in a wide variety of tasks, bus specially on classification. However, understanding their inner mechanisms, especially concerning their interpretability or explainability of their outputs, remains a challenging area of research. Within this field of eXplainable Artificial Intelligence (XAI), there have been proposals of representing neural networks as different models. In this context, the algorithm NN2Poly [1] provides an way of explaining fully-connected feed-forward artificial neural networks, commonly known as multilayer perceptrons (MLPs), using an explicit polynomial representation that only relies on the trained weights of that network and its activation functions. In this work we will show how this method can be used to explain classification outputs on tabular datasets, by means of the obtained polynomial coefficients. One polynomial representation is obtained for each output neuron, i.e., for each class. We will also show the trade off between training the neural network with the needed constraints for NN2Poly to work and the computational cost.

Keywords: XAI , neural networks, interpretability, classification.

References

[1] Morala, P., Cifuentes, J. A., Lillo, R. E., Ucar, I.: NN2Poly: A polynomial representation for deep feed-forward artificial neural networks. IEEE Transactions on Neural Networks and Learning Systems. Early Access (2023)

[I] Tuesday 16, 17:40-18:00, Room 3, session: Classification Methods for Large Datasets (A. Grané)

[II] Deparment of Statistics and uc3m-santander Big Data Institute (IBiDat), Universidad Carlos III de Madrid (uc3m), Spain, pablo.morala@uc3m.es

Similarity Search and LLMs: The RAG Revolution[I]

Plenary Talk

MÜLLER-MOLINA, ARNOLDO[II]

United States

An estimated 60% of LLM (large language models) applications use some form of retrieval-augmented generation (RAG). RAG technologies are usually powered by a similarity search engine. In this talk we first provide an overview of similarity search data structures, types of queries and supporting operations. Then, we share examples of similarity search indexes in the context of RAG. Finally, we share some thoughts about the future of similarity search as the knowledge database that will power LLMs for years to come.

Keywords: similarity search, rag, retrieval augmented generation, llm, llms, large language models, nearest neighbor, proximity search.

[I]Thursday 18, 14:00-15:00, Auditorium, session: Plenary Talk 7

[II]University of Chicago, United States, arnoldomuller@gmail.com

Integration of Deep Learning and Marketing Research for Brand Confusion Prediction and Visual ad Analysis[I]

Communication

NAKAYAMA, ATSUHO[II]

Japan

Image recognition technology, including deep learning, has developed rapidly in recent years. The development of these methods has been significant and is becoming deeper and more widespread in practice. Their prevalence is expected to increase rather than decrease in the future, and the impact on marketing operations is expected to be significant. So, the question is how marketing research should use deep learning approaches. This study considers what further research developments can be expected by combining the deep learning approach with traditional marketing research methods and concepts. The ability to leverage the accumulated knowledge of marketing research and the deep learning approach will give us an advantage in practice. Today, in many consumer goods markets, the physical and chemical differences between competing products are diminishing, largely due to the standardizations of products and production techniques. In such a situation, the important role of marketers is to develop and market attractive advertising that reinforces their brand positioning. They must also ensure that their advertising does not reinforce the brand positioning of competing brands. To date, brand confusion experiments have been used to address this problem. Brand confusion experiments test for brand confusion using confusion matrices, which show the frequency with which each brand is guessed when a print ad is displayed, obtained by presenting consumers with print ads of competing brands and asking them to guess which brand is being advertised. This study analyses the visual content of advertisements using a deep learning approach to predict differentiated positioning and brand confusion. The results will show the differentiation and confusion of each brand based on each company's storytelling and will indicate which communication strategies, such as differentiation and imitation strategies, are being used by each company.

Keywords: brand confusion experiments, brand positioning, brand story, deep learning approach.

[I]Monday 15, 17:00-17:20, Room 3, session: Visualization (J. Nienkemper & S. Lubbe)

[II]Tokyo Metropolitan University, Japan, atsuho@tmu.ac.jp

Weighted Consensus Clustering for Unbiased Feature Importance in Random Forests[I]

Communication

NIANG, NDÈYE[II] Ouattara, Mory[III]

France

Ranking the importance of features in Random Forests (RF) has been shown to be biased in the presence of highly correlated features, especially for highdimensionaldata when the number of features is much larger than the sample size. Several methods have been proposed for unbiased ranking. Among them, the Fuzzy Forest (FF) method [1] combines feature clustering and recursive feature elimination random forests (RFE-RF) and provides relatively unbiased rankings. RFE-RF is performed on each block of features leading to the selection of a percentage offeatures that will be kept in each block. Finally, a RF is applied on the selectedvariables. In this work, through simulation studies, we show that applying differentclustering algorithms yields different feature groups of unequal quality and thusdifferent results concerning important variables. This may lead to an issue for thechoice of the feature clustering algorithm. To overcome this issue, we propose to use new weighted consensus clustering method to get an unique partition [2] on which RFE-RF is performed. The experimental results on simulated data as well as real ones show better performances and stability for the recovery of important variables.

Keywords: random forest, feature importance, weighted consensus.

References

[1] Conn, D., Ngun, T., Li, G., Ramirez, C. M. Fuzzy Forests: Extending Random Forest Feature Selection for Correlated, High-Dimensional Data. Journal of Statistical Software, (2019) 91(9), 1–25. https://doi.org/10.18637/jss.v091.i09

[2] Niang Ndèye and Ouattara Mory : Weighted consensus clustering for multiblock data. In : SFC 2019. https://cnam.hal.science/hal-02471611

[I] Tuesday 16, 09:10-09:30, Room 1, session: Clustering, Classification and Discrimination 1

[II] CEDRIC-CNAM, France, ndeye.niang_keita@cnam.fr

[III] same address

Comparing Precursors for Earthquake Prediction in Chile[I]

Communication

NICOLIS, ORIETTA[II] Varini, Elisa[III] Rotoindi, Renata[IV] Campusano, Efraín[V]
Peralta, Billy[VI] Ruggeri, Fabrizio[VII]

Chile

Predicting earthquakes is a complex challenge due to the intricate processes involved, non-linear correlations among seismic events, and the dependence on numerous variables that are often unidentified or unavailable. Various precursors or indicators have been proposed that could potentially provide information about the likelihood of an earthquake, thereby improving prediction accuracy. In this work, we introduce some precursors based on the history of past seismic events and anomalies in geophysical signals. For the first precursors,we employ the q-exponential probability distribution to analyze temporal variations in seismic parameters (e.g., magnitude, spatial location of epicenters) in earthquake sequences in Chile. Bayesian inference is conducted using data from sliding time windows, each containing a fixed number of events and shifting with each new event. We found that the estimated q-index significantly decreases before strong earthquakes and increases sharply afterward, indicating that it can be considered a reliable indicator of the systems' activation state. We also explore other precursors related to anomalies in geophysical signals such as magnetic fields and cosmic rays. Despite their different origins, we propose common statistical techniques based on change points and wavelets to detect potential earthquakes a few days in advance. By validating these methods with major earthquakes in Chile with magnitudes greater than 8.0 Mw, we identified a common pattern that will be used to predict large-magnitude earthquakes. Finally, a comparison of different precursors is realized for improving earthquake prediction.

Keywords: earthquake prediction, seismic precursors, Bayesian inference.

[I] Tuesday 16, 09:10-09:30, Room 3, session: LACSC: Data Science and Computational Statistics: Theory and applications

[II] Andrés Bello University, Chile, orietta.nicolis@unab.cl
[III] CNR-IMATI, Italy, elisa.varini@cnr.it
[IV] CNR-IMATI, Italy, reni@mi.imati.cnr.it
[V] Andrés Bello University, Chile, efrain.campusano@unab.cl
[VI] Andrés Bello University, Chile, billy.peralta@unab.cl
[VII] CNR-IMATI, Italy, fabrizio.ruggeri@cnr.it

On the Vapnik-Chervonenkis Dimension and Learnability of the Hurwicz Decision Criterion[I]

Communication

NUÑEZ, MANUEL[II] Schneider, Mark[III]

United States

We develop a new axiomatic framework to characterize the classical Hurwicz criterion. Our framework is simpler than other characterizations in the literature. We also study the learnability and falsifiability of the Hurwicz axioms. In particular, we compute the Vapnik-Chervonenkis dimension of the class of Hurwicz preferences, show that the Hurwicz class is PAC (probably approximately correct) learnable, provide a lower bound on the sample size required to learn a concept in this class, and provide an efficient polynomial-time algorithm to either learn or falsify a Hurwicz concept based on data.

Keywords: Hurwicz criterion , machine learning, Vapnik-Chervonenkis dimension, learnability of decision theories.

[I]Thursday 18, 08:30-08:50, Room 2, session: Data Science in Economics, Finance and Management

[II]Department of Operations and Information Management, School of Business, University of Connecticut, United States, Manuel.Nunez@uconn.edu

[III]Culverhouse College of Business, University of Alabama, United States, MASchneider4@cba.ua.edu

A Toolbox for Clustering Ordinal Data in the Presence of Missing Values[I]

Communication

ORTEGA MENJIVAR, LENA[II]

Austria

Ordinal response scales and 'Don't know'-options are ubiquitous responseoptions in surveys. As survey results are a common source for separating respondentsinto (consumer) segments, there is a great need for clustering algorithms able tohandle ordinal, and mixed-with-ordinal data with missing values. While there havebeen significant advances in this field in the last years, especially in the ambit ofmodel-based clustering, solutions tend to be tailored towards specific applications,and no general review is known to the authors. In this work, an in-depth investigationof existing implementations is made. Common strategies for handling missing valuesin ordinal clustering include (1) the imputation and down-weighing of missingdistances, (2) the conjecture of cluster memberships for missing observations in bi-or other multiview-clustering methods from their clustering memberships in non-missing dimensions, and (3) including models for missingness patterns in mixturesof ordinal or mixed-type models. As both categorization and quantification arecommon strategies in the handling of ordinal data, common methods for clusteringinterval and categorical data with missing values will also be included (4), andused to benchmark the methods designed specifically for ordinal data. The collectedimplementations are categorized regarding their assumptions towards variable typesand missingness mechanisms, and applied to real data sets. Their performance isevaluated numerically via common cluster indices, as well as content-wise regardingtheir practicality. Thus, the result of this work is a toolbox of clustering algorithmsdealing with ordinal data and missing values, and serves as decision support forselecting methods in future applications.

Keywords: clustering, ordinal data, missing values.

[I] Tuesday 16, 10:40-11:00, Room 1, session: Clustering, Classification and Discrimination 2

[II] BOKU University of Natural Resources and Life Sciences, Department of Landscape, Spatial and Infrastructure Sciences, Institute of Statistics, Austria, lena.ortega-menjivar@boku.ac.at

Analysis of Seawater Nutrient Concentrations to Assess Submarine Groundwater Discharge Along the Catalan Coast (NW Mediterranean): A Compositional Data Analysis Approach[I]

Communication

ORTEGO, MARÍA ISABEL[II]

Spain

Marine communities provide a variety of ecosystem services. The assessment of nutrients is key in order to evaluate its ecological status. We focus our attention on the Mediterranean Sea. It is considered an oligotrophic sea with very low nutrient concentrations and therefore it is important to consider the inputs from land-based water discharges in coastal areas as they may change the nutrient concentrations and composition. The contributions from rivers and marine outfalls have been extensively studied in the literature, but the effects of Submarine Groundwater Discharges (SGD) on marine ecosystems remain uncertain.This work explores the impact of SGD along the Catalan coast (Western Mediterranean). This is a densely populated area where land-sea interactions and the effects of SGD on marine ecosystems may be substantial. The goal of the study is to establish connections between SGD and the quality of the coastal area through the study of a 23-year dataset. The primary focus lies on the investigation of SGD locations through the analysis of inorganic nutrient composition (NO3, NO2, NH4, PO4, and SiO4) and salinity at 70 coastal stations from a Compositional Data Analysis (CoDA) perspective. Other land-based hydrogeological factors, such as the geological nature of the aquifer are also considered. The relationship between nutrient composition and biological indicators using CoDA techniques is also explored. For instance, chlorophyll as an indicator of photosynthetic activity, may serve as a marker for biological responses to nutrient changes induced by SGD.The present approach takes into account the compositional nature of the data, helps the identification of SGD locations and enables the assessment of its ecological impacts. The results of this study have the potential to provide insights that are valuable for the improvement of coastal ecosystem management.

Keywords: marine ecosystem , inorganic nutrient composition, balances, clr-biplot, log-ratio approach.

[I] Wednesday 17, 10:20-10:40, Room 1, session: Clustering, Classification and Discrimination 6 (A. Grané)

[II] Universitat Politécnica de Catalunya-BarcelonaTECH, Spain, ma.isabel.ortego@upc.edu

Mapping Electoral Behavior and Political Competition: A Comparative Analytical Framework for Voter Typologies and Political Discourses[I]

Communication

PANAGIOTIDOU, GEORGIA[II] Chadjipadelis, Theodore[III]

Greece

This study introduces a methodological framework that integrates HierarchicalCluster Analysis (HCA) and Factorial Correspondence Analysis (AFC) for thecomparative analysis of electoral behavior and political competition. Transcendingtraditional approaches in political science research, this framework offers a comprehensivetool for exploring the complex dynamics of voter behavior, with a particularfocus on young voters in Thessaloniki, Greece. Through the analysis of data collectedfrom over 3,000 participants, this research provides an understanding of the factorsinfluencing first-time voters' electoral decisions and their perceptions of democracyand moral values. Unlike conventional methods that often examine electoral behaviorthrough isolated variables, this study employs a multivariate approach, enablinga more in-depth examination of the interactions between various factors such as politicalmobilization, interest, information sources, and demographic characteristics.The "semantic" map, a pivotal output of the methodological framework, facilitates the direct comparison of behavioral patterns across different voter profiles,thereby highlighting the contrasts and similarities within the electoral landscape.The findings reveal significant insights into the evolution of political attitudes andbehaviors among the youth, demonstrating the method's capability to capture theshifting paradigms of political behavior over time. Moreover, the comparative analysisbrings forward political polarization and competition, offering a dynamic viewof the electoral behavior landscape.

Keywords: electoral behavior , comparative methodology, political competition, hierarchical cluster analysis, factorial correspondence analysis.

[I] Wednesday 17, 08:30-08:50, Room 2, session: Data Science 1

[II] Aristotle University of Thessaloniki School of Political Sciences, Greece, gvpanag@polsci.auth.gr

[III] Aristotle University of Thessaloniki School of Political Sciences, Greece, chadji@polsci.auth.gr

Machine Learning-driven COVID-19 Early Triage and Large-scale Testing Strategies Based on the 2021 Costa Rican Actualidades Survey[I]

Communication

PASQUIER, CARLOS[II] Solís, Maikol[III] Vílchez, Vivian[IV]
Núñez-Corrales, Santiago [V]

Costa Rica

The SARS-CoV-2 pandemic emphasized the importance of mass testing for correct data collection and disease control. This study explores the challenges of optimizing testing, in particular with RT-qPCR and its alternatives. We introduce a population-level strategy that uses predictive mechanisms to assess individual contagion risk, considering factors related to the determinants of health. Using the "Actualidades 2021" survey, which sampled 2003 adults, we set classification models, including logistic regression, Random Forest, Gradient Boosting, and XGBoost. With a prevalence of 0.26 in the sample, we adjust the model to explain the outcome of whether the respondent had COVID-19 or not. The model shows sensitivity and specificity values of 0.79 and 0.76, respectively. Through Monte Carlo simulations, we evaluate the economic and epidemiological impacts of various testing strategies such as pooling, retesting and mixing technologies of RT–qPCR, Antigen and RT–LAMP. In the talk we will discuss how these classification systems could help with the Costa Rican health public policies. The study is available at [1].

Keywords: sars-cov-2 mass testing , classification models, determinants of health, health public policies.

References

[1] Pasquier, C., Solís, M., Vilchez, V., & Núñez-Corrales, S. (2024). Machine learning-driven COVID-19 early triage and large-scale testing strategies based on the 2021 Costa Rican Actualidades survey (p. 2024.04.02.24305223). medRxiv. https://doi.org/10.1101/2024.04.02.24305223

[I] Thursday 18, 10:20-10:40, Room 1, session: Machine Learning

[II] Universidad de Costa Rica, Costa Rica, carlos.pasquier@ucr.ac.cr
[III] Universidad de Costa Rica, Costa Rica, maikol.solis@ucr.ac.cr
[IV] Universidad de Costa Rica, Costa Rica, vivian.vilchez@ucr.ac.cr
[V] University of Illinois Urbana-Champaign, United States, nunezco2@illinois.edu

Mixture Multigroup Structural Equation Modeling: Comparing Structural Relations Across Many Groups[I]

Poster presentation

PEREZ ALONSO, ANDRES FELIPE[II] Rosseel, Yves[III] Vermunt, Jeroen[IV]
De Roover, Kim[V]

The Netherlands

Behavioral scientists often examine the relationships between two or more latent variables or constructs (e.g., attitudes, emotions), and Structural Equation Modeling (SEM) is the state-of-the-art for doing so. When comparing these structural relations among many groups, they likely differ across the groups. However, it is equally likely that some groups share the same relations, and that clusters of groups emerge in terms of the relations between the latent variables. For validly comparing the latent variables' relations among groups, it is important to remember that such variables are indirectly measured via questionnaires and that one should evaluate whether this measurement is invariant across the groups (i.e., measurement invariance). In the case of many groups, often at least some parameters differ across the groups. Current clustering methods using SEM (i.e., mixture SEM methods) force all SEM model parameters (i.e., measurement parameters, structural relations, etc.) to be equal within a cluster, thus also capturing similarities and differences in measurement, which are unrelated to the research question. We propose mixture multigroup SEM (MMG-SEM) to obtain a clustering of groups focused entirely on the structural relations by making them cluster-specific, while allowing for the measurement parameters to be (partially) group-specific to account for measurement non-invariance. In this way, MMG-SEM disentangles differences in structural relations from differences in measurement parameters. We present an evaluation of MMG-SEM's performance in terms of recovering the group-clustering and the group- and cluster-specific parameters as well as an evaluation of different approaches to select the number of clusters (e.g., AIC, BIC, etc.).

Keywords: mixture modeling, structural equation modeling, model selection, structural relations,

[I] Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II] Tilburg University, Department of Methodology and Statistics, The Netherlands, a.f.perezalonso@tilburguniversity.edu
[III] Ghent University, Department of Data Analysis, Belgium, yves.rosseel@ugent.be
[IV] Tilburg University, Department of Methodology and Statistics, The Netherlands, j.k.vermunt@tilburguniversity.edu
[V] KU Leuven, Quantitative Psychology and Individual Differences, Belgium, kim.deroover@kuleuven.be

A Spectral Approach to Evaluating VaR Forecasts: Stock Market Evidence from the Subprime Mortgage Crisis, through COVID-19, to the Russo-Ukrainian War[I]

Communication

PIETRZYK, RADOSLAW[II] Malecka, Marta[III]

Poland

We explore the application of spectral methods in risk management as means ofvalidating VaR models. We propose to replace earlier spectral VaR tests with the testbased on the Anderson-Darling statistic. Based on assumptions relevant to VaR failureanalysis, we experimentally prove that the Anderson-Darling spectral test displaysstrong power to reject inaccurate VaR. Its main advantage over the existing methods isthe combination of two features: the lack of tendency to overreject properly predictedVaR and high sensitivity to limited evidence of incorrectness in VaR predictions. Thus,this test may play an important role in times of change in volatility dynamics, such asoutbreaks of financial crises. We confirm this empirically, based on data starting beforethe subprime mortgage crisis, running through the COVID-19 pandemic, until theoutbreak of the RussoUkrainian war. We give a number of examples when this methodrevealed the inaccuracy of VaR predictions not discovered by commonly used tests.We also show that the proposed spectral test never failed at finding the modelsindicated as incorrect by other tests.

Keywords: spectral test , value-at-risk, VaR test, Anderson-Darling statistic, financial crisis.

[I]Thursday 18, 09:10-09:30, Room 2, session: Data Science in Economics, Finance and Management

[II]Wroclaw Univeristy of Economics and Business, Poland, radoslaw.pietrzyk@ue.wroc.pl

[III]Faculty of Economics and Sociology, University of Lodz, Lodz, Poland,marta.malecka@uni.lodz.pl

PASTE-Boost: P-Value Adjusted Selected Tree Ensembles with Gradient Boosted Improvements[I]

Communication

POOLEY, JOSHUA[II] Lausen, Berthold[III] Mahmoud, Osama[IV]

Great Britain

Generating an optimal or a selected tree ensemble is a method that combines the predictability of low-error decision trees to create a higher performance ensemble. This can be combined with p-value adjusted decision trees which use the significance from statistical tests to determine the most optimal feature and data point to split the data. When combined these methods create a P-value Adjusted Selected Tree Ensemble (PASTE). Gradient boosting can then be applied to the results of PASTE, coined as PASTE-boost, which can further improve accuracy, and reduce any overfitting that can arise from decision trees. This method provides comparable results to other established methods, such as Random Forest and XGBoost across a range of classification datasets, and performs especially well on datasets with categorical predictors, such as the TicTacToe dataset where it predicted perfectly over 5 separate distinct samples. The results suggest that PASTE-boosting is a robust and powerful classification method, that can be used as an alternative to other traditional methods. Due to the methodology of growing the decision trees, the interpretability of the PASTE methodology may also be improved compared to black-box models. Further research can include adapting the algorithm to be applicable to regression problems and improvements to interpretability.

Keywords: decision trees, p-values, gradient boosting, ensemble.

References

[1] Breiman, L.: Random Forests. Mach. Learn. 45(1), 5–32 (2001)

[2] 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, KDD '16, pp. 785–794.

[I] Tuesday 16, 10:20-10:40, Room 1, session: Clustering, Classification and Discrimination 2

[II] University of Essex, Department of Mathematical Sciences, Great Britain, jp22747@essex.ac.uk

[III] University of Essex, Department of Mathematical Sciences, Great Britain, blausen@essex.ac.uk

[IV] University of Essex, Department of Mathematical Sciences, Great Britain, o.mahmoud@essex.ac.uk

A Multivariate Approach for Clustering Functional Data in One and Multiple Dimensions[I]

Communication

PULIDO, BELÉN[II]

Spain

Clustering techniques play a crucial role in unsupervised classification, offering a way to organize complex datasets into coherent groups. While the literature extensively covers clustering techniques in multivariate analysis, the landscape changes when it comes to functional data.Functional datasets present a unique challenge due to their infinite-dimensional nature, making traditional clustering methods less straightforward to apply. To tackle this challenge, we propose to transform the initial functional dataset into a multivariate one.In one-dimensional functional datasets, the original definitions of epigraph and hypograph are considered to obtain the new dataset. For multivariate functional data, new versions of these indexes are introduced. By applying the epigraph and hypograph indexes to obtain a multivariate dataset from a functional one, we reduce dimensionality, rendering the new dataset compatible with standard clustering techniques for multivariate data.Validation of our approach, with both simulated and real datasets, underscores its efficacy in revealing meaningful patterns within functional data. This research serves as a conduit between functional and multivariate analysis, offering a practical solution for clustering functional datasets.

Keywords: clustering , functional data analysis, multivariate analysis, epigraph, hypograph.

S

[I]Wednesday 17, 10:00-10:20, Room 1, session: Clustering, Classification and Discrimination 6 (A. Grané)

[II]uc3m-Santander Big Data Institute (IBiDat), Universidad Carlos III de Madrid, Spain, belen.pulido@uc3m.es

Clustering of Human Gut Microbiome Data Using the Finite Mixture of Generalized Dirichlet-Multinomial Models[I]

Communication

QIN, XIAOKE[II]

Canada

The composition of the human gut microbiome has been reported to be associated with health conditions and the pathogenesis of diseases. Since then, the clustering of microbiome has been of increasing interest, aiming to investigate the subgroups within the population, or enterotypes, in which the people share similar compositions of microbiome. The finite mixture of Dirichlet-multinomial distribution has been widely used for cluster analysis but it is limited by the covariance pattern and the neutrality of Dirichlet distribution. In this paper, we propose to use a finite mixture of the generalized Dirichlet-multinomial model (GDM) which allows for a flexible covariance matrix and less need for neutrality of data. Furthermore, we discuss the non-permutation invariance of GDM. Some examples are presented to show the necessity to account for the orders of datasets in the selection of models. Based on these features, a generalized expectation-maximization algorithm is developed to fit the model, and a stepwise process to permute the column is suggested. A series of simulations are conducted to show the performance of the proposed approach. We apply the model to two human microbiome datasets to capture the latent components and show the correlation structures of the data. The potential association between the order of microbial taxa and the keystone species in the human gut is also discussed. Our model provides a novel perspective for the clustering of compositional data and could mine new information from the permutation-variant data.

Keywords: model-based clustering , compositional data, mixture model, microbiome data.

[I]Thursday 18, 08:30-08:50, Room 1, session: Model-Based Clustering 1

[II]Carleton University, Canada, xiaokeqin@cmail.carleton.ca

Innovating the Banking with Machine Learning: Credit Score for MSMEs[I]

Communication

QUIRÓS MUÑOZ, TATIANA[II] Guevara Villalobos, Álvaro[III]

Costa Rica

The use of innovative machine learning techniques has significantly transformedthe banking sector, including the credit origination process. The adoptionof more robust approaches has led to the creation of more accurate classificationmodels. However, this progress has been accompanied by a dilemma in high-levelbanking discussions: the lack of explainability and interpretability in black-box modelsinvolved in credit access decisions. This is particularly critical issue in banking,as customers and regulators expect to have a reasonable understanding of the variablesthat could improve or deteriorate your chances for credit access. In this study,we will discuss a methodological approach for a credit score for micro, small, andmedium-sized enterprises (MSMEs) at a commercial bank in Costa Rica. To addressthe interpretability challenge, we introduce both local and global interpretabilityperspectives, highlighting the GamiNet method as an enhanced example for neuralnetworks. Developed in recent years, GamiNet aims to maintain the robustness ofa generalized feedforward neural network approach with multiple additive subnetworks.Each subnetwork consists of several hidden layers, allowing us to capturemain effects and pairwise interactions. GamiNet is globally interpretable by design,and also demonstrates competitive accuracy compared to other more established machinelearning methods such as Random Forests.We also performvarious sensibilityanalysis on the model to assess its robustness.

Keywords: banking, credit score, black box, neural network, gaminet, random forests.

References

[1] Sudjianto, A. Zhang: Designing inherently interpretable machine learning models, CorporateModel Risk, Wells Fargo, USA, (2021), arXiv:2111.01743v13.

[2] Z.Yang, A. Zhang, and A. Sudjianto: Gami-net: An explainable neural network based on generalizedadditive models with structured interactions, Department of Statistics and Actuarial Science,The University of Hong Kong, Pokfulam Road, Hong Kong, (2021), arXiv:2003.07132v

[3] Z. Zhao, S. Xu, B. Ho Kang, M. Kabir, Y. Liu, and R.Wasinger: Investigation and improvementof multilayer perception neural networks for credit scoring, Expert Systems with Application,(2014), Available: http://dx.doi.org/10.1016/j.eswa.2014.12.0062.

[I] Thursday 18, 10:00-10:20, Room 2, session: Data Science in Economics, Finance and Management 2

[II] Universidad de Costa Rica, Costa Rica, quirostati@gmail.com

[III] Universidad de Costa Rica, Costa Rica, alvaro.guevaravillalobos@ucr.ac.cr

Improving Employee Attrition with Data Analysis and Machine Learning[I]

Communication

RAMÍREZ RODRÍGUEZ, SERGIO[II] Guevara Villalobos, Álvaro[III]

Costa Rica

The current business environment, characterized by intense competitionin globalized markets and the imperative need for innovation to ensure a competitiveadvantage, has underscored the importance of data analytics as a fundamental toolfor optimizing resources, increasing profits, and enhancing performance indicatorsin companies. One crucial indicator for any company is the minimization of employeeattrition, especially in sectors like Business Process Outsourcing (BPOs).High attrition leads to significant increases in recruitment costs, time required tofind replacements, as well as expenses associated with training new employees andthe loss of accumulated knowledge by departing staff, among other negative effects. Therefore, it is essential for companies to mitigate this risk by using data analyticsand machine learning tools to develop predictive models to identify the likelihoodof employee attrition within a specific period.Various models, including XGBoost, Random Forest, Logistic Regression, and Consensus,were evaluated, comparing metrics such as area under the ROC curve, overallaccuracy, sensitivity, and specificity. XGBoost emerged as superior due to its adeptpredictive capacity, leveraging an ensemble approach with decision trees, regularizationtechniques to forestall overfitting, hyperparameter optimization for optimalconfiguration, and scalability for handling large and complex datasets. Additionally,the model's sensitivity was assessed through stress tests on both observations andpredictor variables. Since implementation, the model has yielded millions of dollarsin the aforementioned savings.

Keywords: employee attrition, machine learning, XGboost, random forest, consensus.

References

[1] Chen, T., Guestrin, C. (2016) Xgboost: A scalable tree boosting system. doi:10.1145/2939672.29397852.

[2] Quin, C., et al. (2021) Xgboost optimized by adaptive particle swarm optimization for credit scoring.Mathematical Problems in Engineering, 2021. doi: 10.1155 /2021/66555103.

[3] Wang, C., Deng, C., Wang, S. (2020) Imbalance-xgboost: leveraging weighted and focal losses forbinary label-imbalanced classification with xgboost. doi: 10.1016/j.patrec.2020.05.035

[I] Thursday 18, 10:40-11:00, Room 1, session: Machine Learning

[II] University of Costa Rica, Costa Rica, sramirez1546@gmail.com

[III] University of Costa Rica, Costa Rica, alvaro.guevaravillalobos@ucr.ac.cr

Mixtures of Quantile-based Factor Analyzers[I]

Helga and Wolfgang Gaul Stiftung Award

REDIVO, EDOARDO[II] Viroli, Cinzia[III]

Italy

In recent years large attention in statistics has focused on dimensionally reduced model-based clustering methods, such as Mixtures of Factor Analyzers, which simultaneously cluster and reduce dimensions using latent variables. Unlike the classical formulation based on Gaussian factors, several adaptations have been made to model complex, non-Gaussian, and diverse data [2, 3]. This work employs quantile-based distributions [1] to model latent variables within factor models, assuming conditional independence among factors. While offering a flexible and parsimonious parameterization, the use of quantile functions increases computational costs, particularly in Bayesian inference. We explore Bayesian estimation of flattened generalized logistic distributions, and then we generalize the model to a mixtures of factor models, providing insights into complex and heterogeneous data structures.

Keywords: factor model, model-based clustering, flattened generalized logistic distribution

References

[1] Gilchrist W (2000) Statistical Modelling with Quantile Functions. Taylor & Francis, Andover, England, UK

[2] Lee S.X., Lin T-I, McLachlan GJ (2021) Mixtures of factor analyzers with scale mixtures of fundamental skew normal distributions. Adv Data Anal Classif 15:481-512. https://doi.org/10.1007/s11634-020-00420-9

[3] Murray P.M., Browne R.P., McNicholas P.D. (2014) Mixtures of skew-t factor analyzers. Computational Statistics & Data Analysis 77:326-335. https://doi.org/10.1016/j.csda.2014.03.012

[I] Wednesday 17, 10:00-10:20, Room 3, session: LACSC session 4: Symbolic Data Analysis

[II] University of Bologna, Italy, edoardo.redivo@unibo.it

[III] University of Bologna, Italy, cinzia.viroli@unibo.it

Riemannian Statistics for Any Type of Data[I]

Communication

RODRÍGUEZ ROJAS, OLDERMAR[II]

Costa Rica

This paper introduces a novel approach to statistics and data analysis, departing from the conventional assumption of data residing in Euclidean space to consider a Riemannian Manifold. The challenge lies in the absence of vector space operations on such manifolds. Pennec X. et al. in their book Riemannian Geometric Statistics in Medical Image Analysis proposed analyzing data on Riemannian manifolds through geometry, this approach is effective with structured data like medical images, where the intrinsic manifold structure is apparent. Yet, its applicability to general data lacking implicit local distance notions is limited. We propose a solution to generalize Riemannian statistics for any type of data.

Keywords: statistics, geometry, manufolds.

[I] Tuesday 16, 14:30-14:50, Room 1, session: Data analysis

[II] Universidad de Costa Rica, Costa Rica, oldemar.rodriguez@ucr.ac.cr

New graphical displays for classification[I]

IFCS Medal

ROUSSEEUW, PETER[II]

Belgium

Classification is a major tool of statistics and machine learning. Several classifiers have interesting visualizations of their inner workings. Here we pursue a different goal, which is to visualize the cases being classified, either in training data or in test data. An important aspect is whether a case has been classified to its given class (label) or whether the classifier wants to assign it to a different class. This is reflected in the probability of the alternative class (PAC). A high PAC indicates label bias, i.e. the possibility that the case was mislabeled. The PAC is used to construct a silhouette plot which is similar in spirit to the silhouette plot for cluster analysis. The average silhouette width can be used to compare different classifications of the same dataset. We will also draw quasi residual plots of the PAC versus a data feature, which may lead to more insight in the data. One of these data features is how far each case lies from its given class, yielding so-called class maps. The proposed displays are constructed for discriminant analysis, k-nearest neighbors, support vector machines, CART, random forests, and neural networks. The graphical displays are illustrated and interpreted on data sets containing images, mixed features, and texts. This is joint work with Jakob Raymaekers and Mia Hubert.

Keywords: class map, discriminant analysis, k-nearest neighbors, label bias, neural net, random forest, silhouette plot, support vector machine.

[I]Thursday 18, 11:40-12:40, Auditorium, session: IFCSmedal

[II]KU Leuven, Belgium, peter.rousseeuw@kuleuven.be, peter@rousseeuw.net

Hypothesis Testing of Mean Interval for p-dimensional Interval-valued Data[I]

Communication

ROY, ANURADHA[II] Montes, Fernando[III]

United States

A new parametric hypothesis test of the mean interval for p-dimensional interval-valued (hyper-rectangles) dataset is proposed under the assumption that the lower bound and the upper bound of an interval are two repeated measurements and the p-dimensional lower bounds and p-dimensional upper bounds have the same variance-covariance matrix. An orthogonal transformation is employed to obtain an equivalent hypothesis test of p-dimensional mean interval of interval-valued dataset in terms of a normal p-dimensional vector of mid-points and a log-normal p-dimensional vector of ranges of the p-dimensional interval-valued dataset. The mean vector of the normal data is tested using Hotelling's T square, while testing for the mean vector of the log-normal data is performed via the construction of a generalized pivotal quantity in a Monte Carlo simulation. The performance of the proposed test is illustrated with a real-life example.

Keywords: generalized pivotal quantity, hypothesis test, interval-valued data, multivariate log-normal, orthogonal transformation.

[I]Tuesday 16, 09:10-09:30, Room 2, session: Symbolic Data Analysis 1

[II]The University of Texas at San Antonio, Management Science and Statistics, United States, Anuradha.Roy@utsa.edu

[III]The University of Texas at San Antonio, Management Science and Statistics, United States, Fernando.Montes2@my.utsa.edu

A Comparison of Multivariate Mixed Models and Generalized Estimation Equations Models for Discrimination in Multivariate Longitudinal Data[I]

Communication

SAJOBI, TOLULOPE[II]

Canada

Discriminant analysis procedures have been developed for classification in multivariate longitudinal data, but the development of such procedures for count, binary or mixed types of outcome variables have not received much attention. Researchers have proposed novel longitudinal discriminant analysis (LoDA) methods using multivariate generalized linear mixed effects models (GLMM) and generalized estimation equations (GEE) to address challenges posed by such data. However, a comprehensive comparison of their predictive accuracy in multivariate longitudinal data remains lacking. This study evaluates the predictive accuracy of these model-based classification procedures via a Monte Carlo simulation study under a variety of data analytic conditions, including sample size, between-variable and within-variable correlation, number of measurement occasions, and number and distribution of outcome variables. Simulation results show that LoDA based on multivariate GEE and GLMM classifiers exhibited similar overall accuracy in multivariate longitudinal data with normal or binary outcome variables. However, the GEE procedure resulted in higher average classification accuracy (between 3% and 23% higher) over the GLMM in multivariate longitudinal data with count or mixed types of outcome

Keywords: multivariate longitudinal data, longitudinal discriminant analysis, generalized linear mixed effects models, generalized estimation equations.

[I] Monday 15, 17:00-17:20, Room 1, session: Modeling Multivariate Data (A. Roy)

[II] Unversity of Calgary, Canada, ttsajobi@ucalgary.ca

Modelling Clusters in Network Time Series with an Application to Presidential elections in the USA[I]

Communication

SALNIKOV, DANIEL[II] Nason, Guy[III] Cortina-Borja, Mario[IV]

Great Britain

Network time series are becoming increasingly relevant in the study of dynamic processes characterised by a known or inferred underlying network structure. Generalised Network Autoregressive (GNAR) models provide a parsimonious framework for exploiting the underlying network, even in the high-dimensional setting. We extend the GNAR framework by introducing the *community-*α GNAR model that exploits prior knowledge and/or exogenous variables for identifying and modelling dynamic interactions across communities in the network. We further analyse the dynamics of *Red, Blue* and *Swing* states throughout presidential elections in the USA. Our analysis shows that dynamics differ among the state-wise clusters.

Keywords: time series clustering, generalised network autoregressive (GNAR) process, community interactions, R-Corbit plot.

[I] Tuesday 16, 17:00-17:20, Room 2, session: Time Series Analysis and Pattern Recognition

[II] Imperial College London, University College London, Great Britain, d.salnikov22@imperial.ac.uk

[III] Imperial College London, University College London, Great Britain, g.nason@imperial.ac.uk

[IV] Great Ormond Street Institute of Child Health Imperial College London, Great Britain, m.cortina@ucl.ac.uk

Drift-switching Local Level Models for Time Series Segmentation[I]

Communication

SAMÉ, ALLOU[II]

France

Many real-world problems involve segmenting temporal data into homogeneous regimes in order to extract relevant features. This operation consists in automatically grouping the points of a single series into clusters associated with contiguous or locally contiguous time intervals. In this work, we are mainly interested in discovering segments that reflect changes in a signal derivative, which are generally associated with dynamic phenomena governed by physical laws. This involves revisiting classical approaches based on mixture models or hidden Markov models. The segmentation approach proposed in this paper is thus inspired by the family of structural time series models. It is an extension of the local level model where the first derivative of the trend component is no longer distributed according to a simple Gaussian distribution, but can switch between different Gaussians via a hidden Markov chain. The resulting model structure, with two levels of latent variables (the denoised unknown trend of the series and its hidden states indicating segment membership), belongs to the family of state-space models. A variational Expectation Maximization (VEM) algorithm is proposed for maximum likelihood estimation. The method is implemented on simulated series and also on real-world series from an energy efficiency context.

Keywords: model-based segmentation, time series, structural model, local level model with drift.

[I] Thursday 18, 16:00-17:00, Coffee area, session: Poster session

[II] Université Gustave Eiffel, COSYS-GRETTIA, France, France, allou-badara.same@univ-eiffel.fr

A Clustering Procedure for Three-Way RNA Sequencing Data Using Data Transformations and Matrix-Variate Gaussian Mixture Models[I]

Communication

SCHARL, THERESA[II]

Austria

RNA sequencing of time-course experiments results in three-way count data where the dimensions are the genes, the time points and the biological units. Clustering RNA-seq data allows to extract groups of co-expressed genes over time. After standardisation, the normalised counts of individual genes across time points and biological units have similar properties as compositional data. We propose the following procedure to suitably cluster three-way RNA-seq data: (1) pre-process the RNA-seq data by calculating the normalised expression profiles, (2) transform the data using the additive log ratio transform to map the composition in the D-part Aitchison simplex to a $D-1$–dimensional Euclidean vector, (3) cluster the trans- formed RNA-seq data using matrix-variate Gaussian mixture models and (4) assess the quality of the overall cluster solution and of individual clusters based on cluster separation in the transformed space using density-based silhouette information and on compactness of the cluster in the original space using cluster maps as a suitable visualisation. The proposed procedure is illustrated on RNA-seq data from fission yeast.

Keywords: compositional data, gene expression, model-based clustering.

References

[1] Scharl, T., Grün, B.: A clustering procedure for three-way RNA sequencing data using data transformations and matrix-variate Gaussian mixture models. BMC Bioinform. (2024) doi:10.1186/s12859-024-05717-6

[I] Thursday 18, 08:50-09:10, Room 1, session: Model-Based Clustering 1

[II] BOKU University, Austria, theresa.scharl@boku.ac.at

Finite Mixture Models for an Underlying Beta Distribution with an Application to COVID-19 Data[I]

Communication

SCHILTZ, JANG[II] Noel, Cédric[III]

Luxembourg

Finite Mixture Models in the sense of Nagin ([2]) are fuzzy logic cluster analysis models for time series. Starting from a sample of trajectories, the aim is to detect a number of subgroups of the sample, so that subjects in the same group exhibit quite similar data trajectories, whereas two subjects from two diferent groups have trajectories that difer in some sense. These models have been generalized by Schiltz ([4]) and are part of a larger strand of models that analyze latent evolutions in longitudinal data.We introduce an extension of Nagin's fnite mixture model to underlying Beta distributions and present our R package ([3]) trajeR which allows to calibrate the model. Then, we test the model and illustrate some of the possibilities of trajeR by means of an example with simulated data.In a second part of the paper, we use this model to analyze COVID-19 related data ([1]) during the frst part of the pandemic. We identify a classifcation of the world into fve groups of countries with respect to the evolution of the contamination rate and show that the median population age is the main predictor of group membership. We do however not see any sign of efciency of the sanitary measures taken by the diferent countries against the propagation of the virus.

Keywords: trajectory analysis , finite mixture model, R package, underlying beta distribution, covid 19.

References

[1] Hasell J. et al., A cross-country database of COVID-19 testing, Scientifc Data, 7(2020), pp. 345.

[2] Nagin D.S., Group-Based Modeling of Development, Harvard University Press, 2005.

[3] Noel C., Schiltz J., trajeR, an R package for cluster analysis of time series, WorkingPaper, University of Luxembourg, Luxembourg, 2022.

[4] Schiltz J., A generalization of Nagin's fnite mixture model, In: Stemmler M., VonÂ Eye A. & Wiedermann W. (eds.), Dependent Data in Social Sciences Research, Springer. 2015, pp. 107-126.

[I] Thursday 18, 10:20-10:40, Room 3, session: Functional Data Analysis 2

[II] University of Luxembourg Department of Finance, Luxembourg, jang.schiltz@uni.lu

[III] University of Lorraine Departement of Commercialisation Techniques, France, cedric.noel@univ-lorraine.fr

Model Selection for Linear Regression Under Data Aggregation[I]

Communication

SCHOONEES, PIETER C.[II]

The Netherlands

Aggregating over individuals belonging to different groups is sometimes unavoidable, such as when data from different views are merged. When performing linear regression, aggregation is known to induce a so-called aggregation bias in the ordinary least-squares (OLS) coefficient estimates compared to those obtained without aggregation. The effect of this aggregation bias on common model selection procedures is however poorly understood. Using simulations based on the matrixvariate normal distribution, we discuss the properties of common selection procedures using a variety of metrics when aggregation is applied.

Keywords: aggregation bias, matrixvariate normal, ordinary least-squares, model selection.

[I]Monday 15, 16:20-16:40, Room 3, session: Visualization (J. Nienkemper & S. Lubbe)

[II]Erasmus University Rotterdam, The Netherlands, schoonees@ese.eur.nl

Accounting for the Shutdown Due to the COVID Pandemic in an Analysis of Multivariate Data from a School and Medical Practice-Based Intervention: the West Philadelphia Asthma Care Implementation Study[I]

Communication

SHULTS, JUSTINE[II]

United States

The West Philadelphia Asthma Care Implementation Program (WEPACC) was designed to improve asthma control in low-income children from communities that are disproportionately impacted by asthma and other serious pediatric health conditions. WEPACC is a randomized control trial that used a factorial design to compare usual care to community health worker (CHW) delivered interventions in primary care, school alone, and combined primary care-school settings. However, the delivery of the different components of the WEPACC intervention was severely impacted by the SARS-CoV-2 pandemic. The shutdown of Philadelphia schools in Spring 2020 (roughly halfway through the study period) prevented the study team from delivering the school components as designed, while social distancing measures led to a striking decrease in asthma morbidity. Some participants completed all study visits prior to the shutdown for the pandemic, while others completed some (or all) of their study visits after the shutdown and therefore did not receive the full "dose" of the intervention as planned. In this presentation I will describe the challenges that we faced in the statistical analysis of data from this trial. I will also discuss and demonstrate the application of some recently published recommendations regarding how to properly modify the statistical analysis plan to account for unplanned interruptions in a clinical trial. For this presentation, the focus will be on multivariate outcomes, in keeping with the spirit of the thematic track, Modeling Multivariate Data (organized by Prof. Anuradha Roy).

Keywords: covid, generalized estimating equations, intervention study, multivariate, quasi-least squares regression.

[I] Monday 15, 16:40-17:00, Room 1, session: Modeling Multivariate Data (A. Roy)

[II] Perelman School of Medicine at University of Pennsylvania, Department of Biostatistics, Epidemiology, and Informatics, United States, shultsj@chop.edu

UMAP Projections and the Survival of Empty Space: A Geometric Approach to High-dimensional Data[I]

Communication

SOLÍS, MAIKOL[II] Hernández, Alberto[III]

Costa Rica

In this work, we explore the potential of applying a type of survival of empty space function to a high dimensional dataset after running it through UMAP. In doing so, we get relevant information on the inner geometric structure of the different clusters obtained from the original data set. Our function is built from the geometry of the data set alone. It looks at different resolutions, the alpha shape that will eventually cover the set. Finally, it will compare its area to that of the smallest window containing the data. The window can be the bounding box or the convex-hull of the data. We apply this to a dataset of human activities. The results show that different activities have different internal geometric structures, in particular the walking activities.

Keywords: survival of empty space function, UMAP, alpha shape, CSR process.

[I] Monday 15, 17:25-17:45, Room 3, session: Big Data and High-Dimensional

[II] Universidad de Costa Rica, Escuela de Matemática, Centro de Investigación en Matemática Pura y Aplicada (CIMPA)., Costa Rica, maikol.solis@ucr.ac.cr

[III] Universidad de Costa Rica, Escuela de Matemática, Centro de Investigación en Matemática Pura y Aplicada (CIMPA)., Costa Rica, albertojose.hernandez@ucr.ac.cr

Machine Learning-Based Classification and Prediction to Assess Corrosion Degradation in Mining Pipelines[I]

Communication

SOW, KALIDOU MOUSSA[II] Ghazalli, Nadia[III]

Canada

The issue of pipeline failure has garnered considerable interest from variousresearch communities due to its notable repercussions on the worldwide economy,as well as the risks associated with leaks, explosions, and expensive periods ofdowntime. This paper aims to build a model for classifying and predicting thecorrosion degradation of a pipe used to transport water in mines by the QuebecMetallurgy Center. To this end, two types of models were developed: three binaryclassification models: SVM, RF, and KNN, yielding F1-measurements of 0.968, 0.969, and 0.945 respectively, and a time series model, LSTM, which, with a lossof less than 0.01, was able to predict average variations in pipeline thickness for 63 days.

Keywords: machine learning , classification, prediction, pipeline corrosion.

[I]Monday 15, 17:45-18:05, Room 2, session: Optimization in Classification and Clustering

[II]University of Quebec at Trois Rivieres, Canada, Kalidou.Moussa.Sow@uqtr.ca

[III]University of Quebec at Trois Rivieres, Canada, Nadia.Ghazalli@uqtr.ca

An Efficient Multicore CPU Implementation of the DatabionicSwarm[I]

Communication

STIER, QUIRIN[II]

Germany

We present an efficiency improved framework for an algorithm exploitingswarm intelligence for self-organized clustering. The algorithm is able to cluster numeric data in three computational steps. First, a projection on two dimensions is achieved by defining each datapoint from the dataset as an agent randomly distributed on a polar grid, on which they self-organize themselves iteratively based on scent emission while their moving radius is cooled, finally resulting in local neighborhoods of similar datapoints. The second step computes the Delaunay triangulation of the projected points and weights the graph edges with the distances from the original high dimensional dataspace and computes the shortest paths with the Dijkstra algorithm. The third step applies hierarchical clustering using the shortest paths in the weighted Delaunay graph. The user can decide the number of clusters based on the resulting dendrogram, but also with a landscape visualization technique of the projection visualizing high-dimensional structures on the generalized U-matrix concept. A higher efficiency is achieved with a parallelized vectorization and minimization of number of operations resulting in the full usage of the CPU. The proposed frameworkis showed to accelerate the performance of a previously implemented sequential algorithm by a factor over 20.

Keywords: self-organization , emergence, unsupervised learning, clustering, swarm intelligence.

[I]Wednesday 17, 09:10-09:30, Room 1, session: Clustering, Classification and Discrimination 5 (A. Grané)

[II]University of Marburg, Germany, Quirin_Stier@gmx.de

Gaussian Mixture Models for Changepoint Detection[I]

Communication

SUBEDI, SANJEENA[II] Dang, Utkarsh[III]

Canada

Changepoint detection aims to find abrupt changes in time series data. These changes denote substantial modifications to the process; they can vary from simple changes in location to a change in distribution. Traditional changepoint detection methods often rely on a cost function to assess if a change occurred in a series. Here, changepoint detection in a clustering framework is investigated, and a novel changepoint detection algorithm is developed using a finite mixture of regressions with concomitant variables. Through the introduction of a label correction mechanism, the unstructured cluster labels are treated as ordered and distinct segment labels. Different kinds of change can be captured using a parsimonious family of models. Performance is illustrated on simulated and real data.

Keywords: changepoint detection, clustering framework, mixture of regressions.

[I]Tuesday 16, 15:10-15:30, Room 1, session: Data analysis

[II]Carleton University, Canada, SanjeenaDang@cunet.carleton.ca

[III]Carleton University, Canada, utkarshdang@cunet.carleton.ca

A Gene Selection Method for Classification with Three Classes Using Proportional Overlapping Scores[I]

Communication

SUWANWONG, ANUSA[II] Harrison, Andrew[III] Mahmoud, Osama[IV]

Great Britain

Genomics experiments, such as microarrays, allow measurements of thousands of gene expression levels within individual samples. They play an important role in distinguishing multiple stages or phenotypes of diseases such as cancer. Implementing a classification that solely relies on specific discriminative genes improves a classifier's interpretability and prediction accuracy. A feature selection method for binary classification within genomics experiments, the Proportional Overlapping Scores (POS), has been proposed, and shown to have good prediction accuracy [1, 2]. Here we propose an extension, named 3-class POS (3cPOS), which deals with the feature selection for classification problems with three classes.3cPOS analyses the gene expressions data and derives a score for the overlapacross the three classes taking into account the proportions of overlapped samples. For each feature, we define a representative mask describing the capability of its gene in distinguishing between the target classes. 3cPOS scores, along with the feature masks, are then utilised to select a subset of informative genes for the classification of interest.3cPOS is compared with Kruskal Wallis Test, Least Absolute Shrinkage andSelection Operator, and Minimum Redundancy and Maximum Relevant, on seven benchmark gene expression datasets. The classification accuracy of the Random Forest, K-Nearest Neighbours, Support Vector Machine, and Extreme Gradient Boost classifiers using the subset of features selected by all methods and the full feature set were examined using 20 repetitions of 5-fold cross validation. Our experiments show that 3cPOS provides an outstanding accuracy for the majority of datasets and classification models.

Keywords: feature selection , microarray classification, proportional overlap score.

[I] Tuesday 16, 14:30-14:50, Room 2, session: Advances in supervised classification

[II] School of Mathematics, Statistics and Actuarial Science (SMSAS), University of Essex., Great Britain, as22799@essex.ac.uk

[III] School of Mathematics, Statistics and Actuarial Science (SMSAS), University of Essex., Great Britain, harry@essex.ac.uk

[IV] School of Mathematics, Statistics and Actuarial Science (SMSAS), University of Essex., Great Britain, o.mahmoud@essex.ac.uk

A New Metric to Classify B Cell Lineage Tree[I]

Communication

TAHIRI, NADIA[II] Farnia, Mahsa[III]

Canada

The B cell lineage tree is a visual representation of the various stages of B cell differentiation and maturation. It shows the progression from hematopoietic stem cells to fully functional antibody-producing cells in the immune system. Accurately classifying these cells requires a reliable metric, similar to an evolutionary tree. Our research introduces a systematic approach for comparing B cell lineage trees that take into account important parameters such as tree topology, branch length, and node abundance. This analytical framework facilitates the exploration of lineage changes over time and allows for the comparison of B cell dynamics within clinical contexts. To the best of our knowledge, we were the first to propose a way of processing heterogeneous data in lineage tree clustering. By addressing the complex challenge of comparing multiple B cell lineage trees, our methodology enhances our comprehension of immune system dynamics in disease contexts.

Keywords: b cell lineage tree , Immunoinformatics, generalized branch length distance, clustering.

[I] Monday 15, 16:40-17:00, Room 2, session: Symbolic Data Analysis 3

[II] University of Sherbrooke, Canada, Nadia.Tahiri@USherbrooke.ca

[III] University of Sherbrooke, Canada, Mahsa.Farnia@USherbrooke.ca

phyDBSCAN: Phylogenetic Tree Density-based Spatial Clustering of Applications with Noise and Automatically Estimated Hyperparameters[I]

Communication

TAHIRI, NADIA[II]

Canada

Phylogenetic analyses commonly produce numerous potential tree topologies, necessitating the resolution of conflicts through consensus-building strategies. However, conventional single-tree consensus methods have inherent limitations. In this study, we propose a novel approach utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, applying the clustering algorithm to the ensemble of candidate trees. Our analysis demonstrates the efficacy of DBSCAN in biological data, particularly its ability to handle low-dimensional datasets and withstand outliers and noise. We introduce a refined DBSCAN algorithm, without hyperparameters (i.e., automatically estimated by the algorithm), specifically designed for the Robinson-Foulds distance. This optimized algorithm efficiently achieves the partitioning of a given set of trees into a singular cluster for homogeneous data or multiple clusters for heterogeneous data.

Keywords: bioinformatics, phylogenetic tree, consensus tree, clustering, DBSCAN, Robinson-Foulds distance.

References

[1] Tahiri, N.,Willems, M. & Makarenkov, V. A new fast method for inferring multiple consensustrees using k-medoids. BMC Evolutionary Biology. 18, 1-12 (2018)

[2] Tahiri, N., Fichet, B. & Makarenkov, V. Building alternative consensus trees and supertreesusing k-means and Robinson and Foulds distance. Bioinformatics. 38, 3367-3376 (2022)

[I] Thursday 18, 09:10-09:30, Room 1, session: Model-Based Clustering 1

[II] University of Sherbrooke, Canada, Nadia.Tahiri@USherbrooke.ca

A Fuzzy Clustering Algorithm with Entropy Regularization for Interval-valued Data[I]

Communication

TENORIO DE CARVALHO, FRANCISCO DE ASSIS[II]

Brazil

Interval-valued data are needed to manage either the uncertainty related to measurements, or the variability inherent to the description of complex objects representing group of individuals. We present a new fuzzy c-means type algorithm based on adaptive Euclidean distances with Entropy Regularization for interval-value data. The improvement in comparison with [1] concerns a new automatic weighting scheme for the interval boundaries [2]. Another improvement concerns the intro- duction of entropy regularization. For that aim a regularization term is adjoined to the maximum internal homogeneity criterion [3], that represents the fuzziness in the form of a weighting factor multiplying the contribution of the regularization function to the clustering criterion. The proposed method optimizes an objective function by alternating three steps aiming to compute the fuzzy cluster representatives, the fuzzy partition, as well as relevance weights for the interval-valued variables. Experiments on synthetic and real datasets corroborate the usefulness of the proposed algorithm.

Keywords: fuzzy clustering , interval-valued data, adaptive distances, entropy regularization.

[I]Tuesday 16, 08:50-09:10, Room 2, session: Symbolic Data Analysis 1

[II]Universidade Federal de Pernambuco, Brazil, fatc@cin.ufpe.br

A Robust Approach of the Clusterwise Regression Method for Distributional Data[I]

Communication

VERDE, ROSANNA[II] Borrata, Gianmarco[III] Balzanella, Antonio[IV]

Italy

This work deals with a robustness approach of the clusterwise regression algorithm (Spaeth H., 1979) for distributional data (CRM-D) (Bock and Diaday, 1999). The CRM-D is based on a new regression method for distributional data, which maps density functions in a Hilbert space via a logarithmic transformation of the derived quantile functions (LDQ) (Petersen A., Muller H., 2016). Let us consider an extension for distributional data of the LDQ functions using a functional representation of the data (Ramsay, J. O. & Silverman, B., 2005). The elements of the explicative distributional variable, assumed as LDQ functions are represented as functional data, considering a smoothing B-splines with knots corresponding to the quantiles of the distributions. CRM-D predicts the response variable in K subclass in which the set of objects is partitioned. In accordance with the clusterwise criterion, the partitioning of the set of objects is performed according to the best fit of the local regression models. The main contribution of the present proposal is to reduce the instability of the results due to the greater variability of the lowest and highest quantiles of the distributions. By using a suitable trimmed of the distributional data, more stable results are achieved for the prediction of the response variable of the K clusters of data. The improving of the fitting of the partitioned data to the respective cluster regression models allows to evaluate the performance of the new approach. Preliminary results on real data have confirmed the effectiveness of the proposed method.

Keywords: distributional data, LDQ, regression model for distributional data.

References

[1] Bock H., Diday E.: Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media, (1999).

[2] Petersen A., Muller H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics, Ann. Statist. 44(1), 183-218, (2016).

[3] Ramsay, J. O. & Silverman, B. W. Functional Data Analysis, 2nd Edition, Springer, New York. (2005). Spaeth, H.: Clusterwise Linear Regression. Computing 22 (4), 367-373 (1979).

[I] Tuesday 16, 10:40-11:00, Room 2, session: Symbolic Data Analysis 2

[II] University of Campania "Luigi Vanvitelli" Dep. of Mathematics and Physics, Italy, Rosanna.VERDE@unicampania.it

[III] Univesrity of Naples Federico II, Dep. Social Science, Italy, gianmarco.borrata@unina.it

[IV] University of Campania “Luigi Vanvitelli” Dep. of Mathematics and Physics, Italy, antonio.balzanella@unicampania.it

Classification of Neuroscientific Data under the Probabilistic Principles of Near-perfect Classification[I]

Communication

VIDAL, MARC[II] Leman, Marc[III] Aguilera, Ana M.[IV]

Belgium

Near-perfect classification is a phenomenon that genuinely occurs in the context of functional data analysis, where random objects (such as functions, fields, shapes...) are assumed to belong to an infinite-dimensional space. We argue that the driving mechanism behind this phenomenon is the Feldman-Hájek Theorem for Gaussian measures [1], and the pathway to achieve this level of accuracy is through the eigenanalysis of a kurtosis operator [2] within a particular RKHS geometry [3]. Through this, we uncover the probabilistic underpinnings explaining why kurtosis has been largely associated with bimodality. We present an analytical framework for functional classification built upon smoothed functional principal/independent components estimators based on a kurtosis optimization criteria [4, 5, 3]. Through extensive simulation studies, we demonstrate the effectiveness of our methods in binary classification problems. The current framework of analysis is applied to model electroencephalographic biomarkers for the diagnosis of depressive disorder by mapping the data into the frequency domain and employing suitable smoothed representations of the log-spectral densities.

Keywords: EEG, Hájek Feldman dichotomy, kurtosis, Picard condition, whitening.

References

[1] Berrendero, J.R., Cuevas, A., Torrecilla, J.L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. JASA, 113(523), 1210-1218

[2] Peña, C., Prieto, J., Rendón, C. (2014). Independent components techniques based on kurtosis for functional data analysis. DES - Working Papers. Statistics and Econometrics, ws141006

[3] Vidal, M., Rosso, M., Aguilera, A.M. (2021). Bi-smoothed functional independent component analysis for EEG artifact removal. Mathematics, 9(11), 1243

[4] Vidal, M., & Aguilera, A. M. (2023). Novel whitening approaches in functional settings. Stat, 12(1), e516

[5] Vidal, M., Leman, M., Aguilera, A.M. (2024) Functional independent component analysis by choice of norm: a framework for near-perfect classification. Under review.

[I] Thursday 18, 09:10-09:30, Room 3, session: Functional Data Analysis

[II] IPEM, Universiteit Gent Departmento de Estadística e I. O., Instituto de Matemáticas, Universidad de Granada Max-Planck-Institut für Kognitions- und Neurowissenschaften, Belgium, marc.vidalbadia@ugent.be

[III] IPEM, Universiteit Gent, Belgium, marc.leman@ugent.be

[IV] Departmento de Estadística e I. O., Instituto de Matemáticas, Universidad de Granada, Spain, aaguiler@uge.es

TabText: A Flexible and Contextual Approach to Tabular Data Representation[I]

Communication

VILLALOBOS CARBALLO, KIMBERLY[II]

United States

In collaboration with Hartford HealthCare (HHC), we have developed highly accurate machine learning (ML) models that predict nine inpatient outcomes (e.g. short-term discharges, ICU transfers, mortality, etc.) using tabular data from electronic medical records. Hundreds of medical staff currently use our models, resulting in a significant reduction in patient-average length of stay and projected annual benefits of $55 - 72$ million for HHC. Given this successful implementation, the question arises: how could we extend these tools for the benefit of hospitals with limited resources, small patient populations, and/or non-standardized healthcare records? To address these challenges, we introduce TabText, a systematic framework that leverages Large Language Models to process and extract contextual information from tabular structures, resulting in more complete and flexible data representations. We show that 1) applying our TabText framework enables the generation of high-performing predictive models with minimal data processing, and 2) augmenting tabular data with TabText representations can significantly improve the performance of standard ML models across all nine prediction tasks, especially when trained with small-size datasets.

Keywords: large language models, healthcare analytics, data augmentation.

References

[1] Carballo, K. V., Na, L., Ma, Y., Boussioux, L., Zeng, C., Soenksen, L. R., & Bertsimas, D. (2022). TabText: A Flexible and Contextual Approach to Tabular Data Representation. arXiv preprint arXiv:2206.10381.

[I] Tuesday 16, 14:30-14:50, Room 3, session: Applications

[II] Massachusetts Institute of Technology, United States, kimvc@mit.edu

Forecasting Realized Volatility: Does Anything Beat Linear Models?[I]

Communication

ZEVALLOS, MAURICIO[II] Ricardi Branco, Rafael[III] Rubesam, Alexandre[IV]

Brazil

We evaluate the performance of several linear and nonlinear machine learning (ML) models in forecasting the realized volatility (RV) of ten global stock market indices in the period from January 2000 to December 2021. We train models using a dataset that includes past values of the RV and additional predictors, including lagged returns, implied volatility, macroeconomic and sentiment variables. We compare these models to widely used heterogeneous autoregressive (HAR) models. Our main conclusions are that (i) the additional predictors improve the out-of-sample forecasts at the daily and weekly forecast horizons

Keywords: volatility forecasting, machine learning, realized volatility, model confidence set, value-at-risk.

[I] Tuesday 16, 10:00-10:20, Room 3, session: LACSC session 1: Data Science

[II] University of Campinas, Brazil, amadeus@unicamp.br

[III] DCIDE, Brazil, ra8branco@gmail.com

[IV] IESEG, France, a.rubesam@ieseg.fr

Keywords Index

Authors Index

Contributions Index by Country