DSSV 2026 — Book of Abstracts

C. Agostinelli[1] and P. Filzmoser[2]

[1]Department of Mathematics, University of Trento, Trento, Italy
[2]Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

June 21, 2026

Claudio Agostinelli

Department of Mathematics
University of Trento
Via Sommarive, 14
I-38123 Povo Trento (TN)
Italy

Peter Filzmoser

Institute of Statistics and Mathematical Methods in Economics
TU Wien
Wiedner Hauptstrasse 8-10
1040 Vienna
AUSTRIA

E-mail addresses: claudio.agostinelli@unitn.it
peter.filzmoser@tuwien.ac.at

2000 Mathematics Subject Classification. Primary 62G35, 62F35, 62K25, 93B35.

ISBN: 9788890333026

Title: Datascience, Statistics and Visualisation 2026

Libero Libri - Claudio Agostinelli, Civezzano (TN), Italy

DSSV 2026

Co-chairs
Claudio Agostinelli, University of Trento
Peter Filzmoser, TU Wien

Programme Committee
Peter Filzmoser, TU Wien (Chair)
Eduard Groeller, TU Wien
Sugnet Lubbe, Stellenbosch University
Matey Neykov, Northwestern University
Katrijn Van Deun, University of Tilburg
Anand Vidyashankar, George Mason University
Veronica Vinciotti, University of Trento

Organising Committee
Claudio Agostinelli, University of Trento (Chair)
Giacomo Francisci, University of Trento
Marcus Mayrhofer, TU Wien
Marta Nai Ruscone, University of Genova
Pier Luigi Novi Inverardi, University of Trento
Una Radojicic, TU Wien
Emanuele Taufer, University of Trento
Veronica Vinciotti, University of Trento

Foreword

Data Science, Statistics & Visualisation (DSSV) is an annual conference of the International Association of Statistical Computing aimed at bringing together researchers and practitioners interested in the interplay of statistics, computer science, and visualisation, and to build bridges between these fields for interdisciplinary research. The goal of this forum is to discuss recent progress and emerging ideas in these different disciplines that contribute to data science, statistics, and visualization. The conference welcomes contributions to practical aspects of data science, statistics and visualization, and in particular those which are linking and integrating these subject areas. Presentations should thus be oriented towards a very wide scientific audience, and can cover topics such as machine learning and statistical learning, the visualization and verbalization of data, big data infrastructures and analytics, interactive learning, advanced computing, and other important themes. We encourage informal contacts and discussions among all the participants.

DSSV has been an annual international conference since 2017. The previous DSSV meetings were held in Kruger National Park, South Africa (2025), Fairfax, VA , USA (2024), Antwerp, Belgium (2023), Tainan, Taiwan, ROC (2022), Rotterdam, The Netherlands (2021), Durham, NC, USA (2020), Kyoto, Japan (2019), Vienna, Austria (2018), Lisbon, Portugal (2017).

The contributions to DSSV 2026 are reported in this *Book of Abstracts* either in the form of abstracts or extended abstracts.

The Co–chairs

Claudio Agostinelli

Peter Filzmoser

The conference is organized by

University of Trento
Department of Mathematics
Department of Economics and Management
www.unitn.it

TU Wien
Institute of Statistics and Mathematical Methods in Economics
www.tuwien.at/en/mg/cstat

Centro Internazionale per la Ricerca Matematica
cirm.fbk.eu

International Association for Statistical Computing (IASC)
iasc-isi.org

Scientific Program

29 June 2026 (Monday)

08:00 Registration open

09:00–09:15 **Opening** Room **A101**

09:15–10:15 **Plenary Session**

PS01: *"Optimal Self-Distillation in Ridge Regression: Sharp Asymptotics and One-Shot Tuning"*
Alessandro Rinaldo (The University of Texas at Austin)
Chaired by Peter Filzmoser (TU Wien) Room **A101**

10:15–10:45 Coffee Break

10:45–12:25 **Contributed Sessions**

CS01 – Environmental and Agricultural Applications
Chaired by ***Emanuele Taufer*** *(University of Trento)* Room **A110**

"Seasonal Variability in Thermodynamic Conditions Preceding Heavy Rainfall in the Free State, South Africa"
Ilse Schoeman (North-West University)

"Classification of Highly Colinear Infrared Spectral Data"
Sugnet Lubbe (Stellenbosch University)

"Beyond the Smoke: What Statistics Reveal About Brazilian Wildfires"
Paulo Canas Rodrigues (Federal University of Bahia)

"Analysing Tail Dependencies of Temperature in Agriculturally Active Regions Across South Africa"
Vusi Ntiyiso Masingi (North-West University)

"Analysing the Efficiency of Deseasonalisation Methods in Removing Seasonal Bias in Outliers"
Rofhiwa Netshiomvani (North-West University)

CS02 – Interactive Visualisation and Statistical Software
*Chaired by **Steffen Frey** (University of Groningen)* Room **A101**

"CougarStats: Integrating Computational Statistics and Interactive Visual Analytics"
Ashok Krishnamurthy (Mount Royal University)

"Visualising Education Policy Reforms and Inequality Indicators in Europe since 2000"
Maria Symeonaki (Panteion University)

"Interactive Visual Parameter Exploration in Functional Data Analysis"
Truls H. Jakobsen (University of Bergen)

"When the Model Won't Invert: Interactive Visual Optimization of Complex Simulations"
Kresimir Matkovic (VRVis Research Center, Vienna)

"Interactive Calibrated-Axes Biplots in R with the bipl5 Package"
Ruan Buys (Stellenbosch University)

CS03 – Clustering and Methods for Clustered Data
*Chaired by **Patrick Groenen** (Erasmus University)* Room **A102**

"Sparse Feature Group K-Means"
Ndèye Niang (Conservatoire national des arts et métiers)

"Robust Clustering for Matrix-Variate Data"
Marcus Mayrhofer (TU Wien)

"The Bixplot: A Variation on the Boxplot Suited for Bimodal Data"
Peter Rousseeuw (University of Leuven)

"Sparse Weighted Multi-View Clustering"
Ndèye Niang (Conservatoire national des arts et métiers)

"Multivariate Functional Mahalanobis Distance with Application to Clustering"
Una Radojicic (TU Wien)

12:25–14:00 Lunch

14:00–15:30 **Invited Sessions**

IS01 – Dynamic Network Modelling
Chaired by ***Veronica Vinciotti*** *(University of Trento)* Room **A101**

"Modelling and Inference of Relational Events"
Ernst-Jan Camiel Wit (Università della Svizzera italiana)

"Identifying Model Misspecification in Stochastic-Actor Oriented Models"
Viviana Amati (University of Milano-Bicocca)

"Beyond Binary Ties: Modeling the Evolution of Relational States"
Rūta Juozaitienė (Vytautas Magnus University)

IS02 – Dependence in Multivariate Settings
Chaired by ***Emanuele Taufer*** *(University of Trento)* Room **A102**

"Selection Confidence Sets for Equally Weighted Portfolios"
Sandra Paterlini (University of Trento)

"Graphical LASSO for Estimating Associations in Elliptically Symmetric Distributions: An EM-Based Approach"
Abdaljbbar Dawod (University of Debrecen)

"Shrinkage and Visualization of Multivariate Gram-Charlier Approximations"
Gyorgy Terdik (University of Debrecen)

15:30–16:00 Poster Session and Coffee Break

"The Colony as a Network: Graph-Theoretic Morphometrics in *Eunicella cavolini* (Cnidaria: Octocorallia)"
Pietro Apolloni (University of Trento)

"CRUSH: Causal Regularization Under Shifted Heterogeneity"
Alessia Berarducci (Università della Svizzera Italiana)

"A Modified Two One-Sided Test for Reference-Scaled Average Bioequivalence"
Chieh Chiang (Tamkang University)

"Research on Data Assets Pricing Based on LLM Agent"
Haibo Han (Lanzhou University of Finance and Economics)

"Centralised Validation and AI Report Generation: Accelerating Statistical Dissemination at an NSO"
Akbr Kanyesigye (Uganda Bureau of Statistics)

"Estimating Poverty Incidence in Côte d'Ivoire Using Satellite Imagery"
Kinin Mamadou Kone (National Statistics Agency (ANStat), Cote d'Ivoire)

"Fine Motor Performance and Global Cognitive Function with Incomplete Data in the Population-Based CHRIS Study"
Roberto Melotti (Eurac Research)

"Contrasting Disinformation: A Mixture of Hawkes Processes for the Identification of Coordinated Inauthentic Behaviour"
Elisa Muratore (University of Trento)

"A Compositional Data Analysis Framework for Diagnosing LLM Reasoning over Time Series Anomalies"
Ceylan Yozgatlıgil (Middle East Technical University)

16:00–17:30 Invited / Organized Sessions

IS03 – Scalable Likelihood-Based Inference for Complex Statistical Models
Chaired by ***Davide Ferrari*** *(Free University of Bozen-Bolzano)* Room **A101**

"Scalable Inference for Individual-Based Epidemic Models"
Lorenzo Rimella (University of Bergamo)

"Pairwise Fisher Transformation of a Conditional Correlation Matrix"
Marco Plazzogna (University of Geneva)

"Sparse Unbiased Estimating Equations via Likelihood Score Alignment"
Giulia Bertagnolli (Free University of Bozen-Bolzano)

OS01 – Modern AI/ML for Clinical Evidence and Decision Support
Chaired by ***Liangyuan Hu*** *(Rutgers University)* Room **A102**

"Bayesian Nonparametric Inference for Dynamic Treatment Strategies with Missing Time-Varying Covariates"
Jason Roy (Rutgers University)

"Advancing Clinical Research with Large-Scale EHR Data and Machine Learning"
Yi Guo (University of Florida)

"The Eidos in the Echo: Finding a Blessing in LLM Hallucination"
Sijian Wang (Rutgers University)

"Predicting Expert MRI Usability Decisions from Acquisition Metadata Alone: A Calibrated, Interpretable Model in ADNI3"
Tia Warrick (Juniata College)

OS02 – Advances in Mixture Modeling and Model-Based Clustering
Chaired by ***Yana Melnykov*** *(University of Alabama)* Room **A110**

"Latent Covariance Clustering for Few-Shot Classification in Trace Data Analysis"
Semhar Michael (South Dakota State University)

"A State-Restricted Hidden Markov Model for Authorship Attribution of the Deutero-Pauline and Pastoral Epistles"
Shuchismita Sarkar (Bowling Green State University)

"On Fuzzy Classification Indices and Their Variability Assessment"
Volodymyr Melnykov (University of Alabama)

"transDA: An R Package for Transformation Discriminant Analysis"
Yana Melnykov (University of Alabama)

20:00–22:00 Welcome Party

30 June 2026 (Tuesday)

08:00 Registration open

09:00–10:00 Plenary Session

PS02: *"Statistical Modeling of High-Order Interactions: Recent Developments and Future Challenges"*
Catherine Matias (Sorbonne University)
Chaired by Veronica Vinciotti (University of Trento) Room **A101**

10:00–10:30 JDSSV Award Session

JA01 Room **A101**

"Cellwise Outliers: From Identification to Regression"
Jakob Raymaekers (University of Antwerp)

Patrick Groenen (Erasmus University)

10:30–11:00 Coffee Break

11:00–12:40 Contributed Sessions

CS04 – Causal Inference, Graphical and Network Models
Chaired by ***Veronica Vinciotti*** *(University of Trento)* Room **A101**

"Hyperevent Network Modelling of Partially Observed Gossip Data"
Veronica Poda (University of Trento)

"Beyond Pairwise: Investigating Higher-Order Statistical Behaviour in Network Psychometrics"
Niels Van Santen (Ghent University)

"Age-Specific Mortality in Italian Provinces via Functional Generalized Linear Mixed Models"
Matteo Scianna (University of Trento)

"Bias-Variance Propagation from Precision Matrix Estimation to Minimum-Variance Portfolios"
Alessandro Fulci (University of Trento)

"Sensitivity Analysis for Causal Effects Measured on the Odds Ratio Scale"
Elena Stanghellini (University of Perugia)

CS05 – Computational and Algorithmic Advances
Chaired by ***Gyorgy Terdik*** *(University of Debrecen)* Room **A102**

"Efficient and Scalable Bayesian Inference for Joint Modelling Longitudinal and Event History Data"
Guangquan Li (Northumbria University)

"Recursive Computation of Multivariate Hermite–Gaussian Integrals with Applications"
Emanuele Taufer (University of Trento)

"Scalable Estimation of Large Generalized Additive Models"
Claudia Collarin (University of Edinburgh)

"On Stochastic Structure of Thinning-Based Non-Negative Vector-Valued ARMA Processes"
Márton Ispány (University of Debrecen)

"interlace: Closing the Mixed-Effects Modelling Gap in Python"
Helio Pais (The StepStone Group)

CS06 – Distance and Classification
Chaired by ***Una Radojicic*** *(TU Wien)* Room **A110**

"A Supervised Distance Metric for K-Nearest Neighbors with Mixed-Type Data"
Carlo Cavicchia (Erasmus University)

"Depth Functions for Tree-Indexed Data"
Giacomo Francisci (University of Trento)

"A Robust Distance Concept for Random Surfaces with Application to Classification"
Leopold Micheler (TU Wien)

"Cellwise Robust Gaussian Mixture Model for Multi-Group Data with Label Noise"
Patricia Puchhammer (TU Wien)

"Robust Data Cleaning for Tensor-Valued Observations, with Application to GCxGC-MS Data"
Peter Filzmoser (TU Wien)

12:40–14:00 Lunch

14:00–15:30 Invited Sessions

IS04 – Modern Causal Inference
Chaired by ***Matey Neykov*** *(Northwestern University)* Room **A101**

"Addressing an Extreme Positivity Violation to Distinguish the Causal Effects of Surgery and Anesthesia via Separable Effects"
Caleb Miles (Columbia University)

"Polynomial-Time Near-Optimal Estimation over Certain Type-2 Convex Bodies"
Matey Neykov (Northwestern University)

"Causal Invariance in Graphical Models with Latent Variables"
Marco Borriero (University of Florence)

IS05 – Visualization Meets Data Science and Statistics
Chaired by ***Eduard Gröller*** *(TU Wien)* Room **A102**

"MLMC: Visualizing Multi-Label Classification"
Torsten Möller (University of Vienna)

"Mathematical Optimization for Scalable Scientific Visualization"
Steffen Frey (University of Groningen)

"On the Role of Interaction in Visual Data Science"
Helwig Hauser (University of Bergen)

15:30–16:00 Poster Session and Coffee Break

"The Colony as a Network: Graph-Theoretic Morphometrics in *Eunicella cavolini* (Cnidaria: Octocorallia)"
Pietro Apolloni (University of Trento)

"CRUSH: Causal Regularization Under Shifted Heterogeneity"
Alessia Berarducci (Università della Svizzera Italiana)

"A Modified Two One-Sided Test for Reference-Scaled Average Bioequivalence"
Chieh Chiang (Tamkang University)

"Research on Data Assets Pricing Based on LLM Agent"
Haibo Han (Lanzhou University of Finance and Economics)

"Centralised Validation and AI Report Generation: Accelerating Statistical Dissemination at an NSO"
Akbr Kanyesigye (Uganda Bureau of Statistics)

"Estimating Poverty Incidence in Côte d'Ivoire Using Satellite Imagery"
Kinin Mamadou Kone (National Statistics Agency (ANStat), Cote d'Ivoire)

"Fine Motor Performance and Global Cognitive Function with Incomplete Data in the Population-Based CHRIS Study"
Roberto Melotti (Eurac Research)

"Contrasting Disinformation: A Mixture of Hawkes Processes for the Identification of Coordinated Inauthentic Behaviour"
Elisa Muratore (University of Trento)

"A Compositional Data Analysis Framework for Diagnosing LLM Reasoning over Time Series Anomalies"
Ceylan Yozgatlıgil (Middle East Technical University)

16:00–17:30 **Invited Sessions**

IS06 – Advances in Latent Variable Modelling
Chaired by ***Katrijn Van Deun*** *(Tilburg University)* Room **A101**

"Robust Estimation of Structural Equation Models with Ordinal Data"
Andreas Alfons (Erasmus University Rotterdam)

"Structural Equation Modeling with Instrumental Variables"
Yves Rosseel (Ghent University)

"Unifying Weights and Loadings in Sparse Component Analysis via Equality and Cardinality Constraints"
Hetvi Chaniyara (Eindhoven University of Technology)

IS07 – AI, Statistics, and Visualization
Chaired by ***Anand N. Vidyashankar*** *(George Mason University)* Room **A102**

"Interpretable Frequency Band Summary Measures and Analysis for Multiple Nonstationary Biomedical Time Series"
Scott Bruce (Texas A&M University)

"Robust Diffusion Models via Divergence-Induced Weighted Denoising"
Lei Li (LunarAI LLC)

"Hellinger-Type Losses for Generative Adversarial Networks"
Giovanni Saraceno (University of Padova)

19:30–22:00 Social Dinner

01 July 2026 (Wednesday)

08:30 Registration open

09:00–10:00 **Plenary Session**

PS03: *"Humans, Machines, and the Space Between: Engineering Effective Human-AI Partnerships"*
Menna El-Assady (ETH Zurich)
Chaired by Helwig Hauser (University of Bergen) Room **A101**

10:00–10:30 Coffee Break

10:30–11:10 Contributed Sessions

CS07 – Dimension Reduction Methods
*Chaired by **Patricia Puchhammer** (TU Wien)* Room **A101**

"Principal Component Analysis for Interval-Valued Data with Multiple Number of Observations from Each Subject"
Anuradha Roy (The University of Texas at San Antonio)

"Globally Aligned Principal Component Analysis for Multi-Group Data"
Hedayat Fathi (Université Laval)

CS08 – Microbiome Data Analysis
*Chaired by **Sugnet Lubbe** (Stellenbosch University)* Room **A102**

"A Robust and Flexible Nonparametric Approach for Longitudinal Microbiome Data Analysis"
Taesung Park (Seoul National University)

"Compositional Microbiome-Based Signatures Associate with General Health Status"
Meritxell Pujolassos (University of Vic)

CS09 – Bayesian Model Selection
*Chaired by **Luca Greco** (University Giustino Fortunato)* Room **A110**

"Bayesian Variable Selection in Generalized Linear Models"
Lucia Filippozzi (University of Trento)

"Contrasting Disinformation: A Mixture of Hawkes Processes for the Identification of Coordinated Inauthentic Behaviour"
Elisa Muratore (University of Trento)

11:10–12:40 Invited Sessions

IS08 – Some Data Science, Some Statistics and Some Visualisation
Chaired by ***Sugnet Lubbe*** *(Stellenbosch University)* Room **A101**

"Data Science and Generative-AI"
Humphrey Brydon (University of the Western Cape)

"Outlier Detection in Histogram-Valued Data"
Paula Brito (University of Porto)

"Interpretable Kernels"
Patrick Groenen (Erasmus University)

IS09 – Complex Modelling and Robustness
Chaired by ***Claudio Agostinelli*** *(University of Trento)* Room **A102**

"Multivariate Wrapped Normal Estimation with Missing Values"
Luca Greco (University Giustino Fortunato Benevento)

"Divergence-Based Methods for Diffusion LLMs: Semantic Prompt Shifts, Unmasking, and Robustness"
Anand N. Vidyashankar (George Mason University)

"Copula-Based Models for Spatially Dependent Cylindrical Data"
Francesca Labanca (University of Florence)

12:40–12:50 Closing

12:50–14:00 Lunch

Abstracts

Robust estimation of structural equation models with ordinal data

M. Welz[a], P. Mair[b] and A. Alfons[c]

[a] *University of Zurich,* [b] *Harvard University,* [c] *Erasmus University Rotterdam*

Structural Equation Models (SEMs) are typically fitted to a given correlation matrix, which is commonly estimated from a sample of Likert-type rating data. However, noisy or low-quality observations—such as (but not limited to) careless responses [1]—might be present in the data, which can introduce a sizable bias in correlation estimates [2, 3]. We demonstrate that this bias is inherited by the SEM estimate, possibly leading to worse model fit and biased estimates of factor structure. As a remedy, we propose to use a robust estimate of a polychoric correlation matrix [3]. We show through simulation studies and empirical applications that fitting a SEM to a robustly estimated polychoric correlation matrix can substantially improve SEM fit, enhance the accuracy of parameter estimates, and help identify potentially low-quality responses. In particular, we demonstrate how the fit of commonly used SEM estimators such as maximum likelihood or least-squares-based approaches like diagonally weighted least squares (DWLS) can be improved by using a robustly estimated polychoric correlation matrix. Our proposed procedure is implemented in the free open-source R package robcat, which is implemented using fast and efficient C++ code.

Keywords: Careless responding, Polychoric correlation, Partial misspecification

References

[1] A. Alfons and M. Welz (2024). Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace? *Social and Personality Psychology Compass*, **18**(2), e12941.

[2] M. Welz, A. Archimbaud, and A. Alfons (2024). How much carelessness is too much? Quantifying the impact of careless responding. *PsyArXiv*.

[3] M. Welz, P. Mair, and A. Alfons (2025). Robust estimation of polychoric correlation. *Psychometrika*, published online.

Smarter Jittering for Scatterplot

Natalia da Silva[a], Ignacio Alvarez-Castro[a], Dianne Cook[b] and Jayani P., Gamage[b].

[a] *Instituto de Estadística, Universidad de la República, Uruguay.*
[b] *Department of Econometrics and Business Statistics, Monash University, Australia*

Overlapping points in scatterplots can hide important patterns, making it hard to see how data are distributed or related especially when values are discrete or take only a few levels. Jittering is often used to spread points apart, but most existing methods simply add small, independent noise to each axis.

We introduce a new approach to jittering in two dimensions that spreads overlapping points more effectively while still reflecting the structure of the data. The method allows different ways of generating noise such as uniform, Gaussian, or quasi-random (Sobolev) patterns and can also control the direction and alignment of the spread. This can be applied either across the whole dataset or locally, based on nearby points.

These methods are implemented in the `ggscatite` R package (available in https://github.com/natydasilva/ggscatite), which extends `geom_jitter()` to support coordinated two-dimensional jittering. This gives users more control over how points are displayed and makes dense scatterplots easier to interpret.

Keywords: Scatter plot, jittering, quasi-random noise

The colony as a Network: Graph-Theoretic Morphometrics in *Eunicella cavolini* (Cnidaria: Octocorallia)

B. Coda[a] and P. Apolloni[b]

[a]*Kanazawa University,* [b]*University of Trento*

Eunicella cavolini is one of the most ecologically prominent gorgonian corals in the Mediterranean, growing in intricate branching shapes that usually encode structural information at multiple levels. Despite this complexity, extracting topologically meaningful descriptors using quantitative frameworks is not common practice. This study addresses the following research question: does *Eunicella cavolini* express a conserved colony-level architectural blueprint quantifiable through terminal branch fraction and branch length distribution structure?

We present a morphometric pipeline applied to eleven colonies sampled across three sites along the coast of Barcelona, Spain. We acknowledge that this sample size limits statistical generalization; knowing this, results should be interpreted as hypothesis-generating rather than confirmatory. Non-destructive field measurements, such as total colony height, width, and individual branch lengths, were used as inputs throughout.

The analytical pipeline operates at two levels. First, Strahler stream ordering quantifies architectural self-similarity within each colony. Second, each colony is modeled as a weighted arborescence, a directed tree graph, in which nodes represent branching junctions and tips, and edge weights correspond to measured branch lengths. From this representation, we derive degree distribution, mean path length and betweenness centrality. This last metric is of particular functional interest: branches with high centrality constitute structural bottlenecks whose removal would fragment a great portion of the colony, providing a proxy for mechanical and ecological vulnerability.

Preliminary results reveal that the proportion of terminal leaf nodes is at 73–74% of total segments across individuals, suggesting a species-level topological constraint. Additionally, branch length distributions maintain a consistent relative structure, with segment counts showing an approximately constant ratio at the most peripheral tier. Both findings require validation

on a larger, environmentally stratified dataset. Nevertheless, they suggest that *E. cavolini* expresses a conserved architectural blueprint at both the topological and morphometric levels.

Keywords: *Eunicella cavolini*, Gorgonian Network Architecture, Gorgonian Morphometrics.

CRUSH: Causal Regularization Under Shifted Heterogeneity

A. Berarducci[a], M. Lembo[a], V. Vinciotti[b] and E.C. Wit[a]
[a] *USI Università della Svizzera Italiana,* [b] *University of Trento,*

Keywords: Causal Inference, Invariant Causal Prediction, Out of Sample Risk Minimization.

Prediction is one of the most important uses of statistical methods. However, predictive systems typically struggle when the environment in which they were trained changes. The central research question is to identify and estimate stable models when only arbitrary learning environments are available. Existing approaches, such as Causal Dantzig [2] and Causal Regularization [1], directly identify the causal parameter as the unique solution whose residual moments remain equal across environments. However, when perturbations are weak or the shift-induced difference matrix is ill-conditioned, this may lead to unstable estimates. This work extends Causal Regularization, obtaining risk guarantees under heterogeneous shifts with few assumptions. Causal Regularization Under Shifted Heterogeneity (CRUSH) is a generalization of Causal Regularization that operates with two shifted environments, without assuming that either is unperturbed. In general, the risk difference becomes non-convex and may exhibit saddle-point geometry. We show that the causal parameter β_{CP} is characterized as a stationary point of the generalized risk difference. Furthermore, defining a sieve of out-of-sample distributions, we derive a worst-risk decomposition in each of these increasing sets of data environments with a closed-form solution for $\hat{\beta}$ that interpolates between the OLS and the causal parameter while guaranteeing uniqueness even under non-aligned shifts. Our results show that even though the estimator was trained on a limited set of in-sample data environments, we can obtain prediction stability in a large set of perturbed or non-stationary data environments with minimal assumptions on the shifts.

References

[1] Lucas Kania and Ernst Wit. Causal Regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees, 2025.

[2] Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions, 2018.

Sparse Unbiased Estimating Equations via Likelihood Score Alignment

G. Bertagnolli[a], Z. Huang[b] and D. Ferrari[a]

[a] *Free University of Bozen-Bolzano, IT,* [b] *RMIT University, Melbourne, AU*

Estimating equations are central to statistical inference, underpinning likelihood- and moment-based methods. Consider a parametric model $\mathcal{M} = f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p$ and a sample $\left(Y^{(1)}, \ldots, Y^{(n)}\right)$. We study inference for θ in high-dimensional regimes ($p \geq n$), based on an unbiased estimating function $S : \Theta \times \mathcal{Y} \to \mathbb{R}^p$ satisfying $\mathbb{E}\theta[S(\theta, Y)] = 0$ and the associated system $\sum i = 1^n S(\theta, Y^{(i)}) = 0$. Such functions include moment conditions and composite likelihood scores. In high dimensions, the system is typically ill-conditioned, motivating a sparsity assumption: the true parameter θ_0 has support $A_0 = j : \theta_{0j} \neq 0$ with $|A_0| \ll p$. Building on optimal estimating function theory [1, 2], we look for an optimal estimating function $\tilde{S} = W_0 S$ in the class of linear transformations of S, where W_0 is the minimiser of the convex criterion

$$\mathcal{L}_\lambda(W) = \frac{1}{2} \operatorname{tr}\left(W \Sigma_S W^\top\right) - \operatorname{tr}\left(H_S W^\top\right) + \operatorname{pen}_\lambda(W).$$

with Σ_S, H_S being the variability and sensitivity matrices of S respectively. The penality $\operatorname{pen}_\lambda(W)$ enforces sparsity directly at the estimating function level, also providing a link between the sparsity pattern of W and the active set A_0. Under standard conditions for penalised models, we establish selection consistency and asymptotic normality on the estimated active set. We also propose a blockwise proximal scoring algorithm for efficient computation of $(\widehat{W}, \hat{\theta})$, and illustrate the method on problems including linear regression and sparse multinomial models.

Keywords: High-dimensional statistics, Estimating equations, Sparsity

References

[1] Vidyadhar P Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

[2] Christopher C Heyde. *Quasi-likelihood and its application: a general approach to optimal parameter estimation.* Springer, 1997.

Causal invariance in graphical models with latent variables

M. Borriero[a], M. Lupparelli[a], G. M. Marchetti[a] and V. Vinciotti[b]

[a]*Department of Statistics, Computer Science and Applications, University of Florence,* [b]*Department of Mathematics, University of Trento*

The purpose of causal discovery is to identify causal relationships among variables from observational or interventional data, typically represented by a directed acyclic graph (DAG). In [1] the authors introduced the invariant causal prediction methodology that enables the identification of the causal parents of target variables by exploiting the stability of causal effects across different experimental settings. However, when some parents are unobserved, the induced graph over the observed variables may no longer be a DAG. In fact, in general it belongs to a broader class of graphs, and moreover it may not be unique, complicating causal inference. We examine in detail relevant configurations of latent parents, with particular attention to the case of hidden confounders, we characterize the induced graph and formalize the conditions under which causal invariance is preserved for identification of the observed parents. Necessary and sufficient conditions for testing such invariance are formally established for a (multivariate) Gaussian target.

Keywords: acyclic directed mixed graphs; identifiability; mediation analysis.

References

[1] J. Peters, P. Bühlmann and N. Meinshausen (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78**, 947–1012.

Outlier Detection in Histogram-valued Data

Ana Martins[a], Sónia Dias[b], Paula Brito[c], and Peter Filzmoser[d]

[a]*Institute of Electronics and Informatics Engineering of Aveiro, Aveiro, Portugal*
[b]*Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal*
[c]*Fac. Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal*
[d]*Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria*

We introduce a novel method for multivariate outlier detection in histogram-valued data. The proposed method is based on Donoho's outlyingness measure, and is inspired by the approach of [1] for functional data, where observations are projected into a one-dimensional space. In our case, this projection takes advantage of the linear combination proposed in [2] based on the representation of empirical distributions by their quantile functions. Assuming a Uniform distribution within each sub-interval of the observed histograms, these quantile functions are piecewise linear functions. The proposed outlyingness measure may rely on either the L1-Wasserstein or on the Mallows (L2-Wasserstein) distance to compare distributions.

An extensive simulation study, considering data with different distributions, and different outlier proportions and severity, shows that the proposed approach is efficient in detecting atypical observations, even in cases where they are close to the regular ones. The method is further applied to two real datasets, regarding flight data and Austrian meteorological stations data, allowing to identify atypical cases.

Keywords: Distributional data, Outlyingness measure, Symbolic Data Analysis

References

[1] Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, **24**(2):177–202.

[2] Dias, S. and Brito, P. (2015). Linear regression model with histogram-

valued variables. Statistical Analysis and Data Mining: The ASA Data Science Journal, **8**(2):75–113

Interpretable Frequency Band Summary Measures and Analysis For Multiple Nonstationary Biomedical Time Series

C. Brubaker, R. Lee and S. A. Bruce

Texas A&M University

This work develops a likelihood-based methodology for estimating frequency bands in collections of nonstationary time series that exhibit replicate-specific spectral variability, with a motivating application to pupil diameter dynamics in children with and without attention deficit hyperactivity disorder (ADHD). The proposed framework approximates time-varying spectra using piecewise-smooth functional summaries over data-adaptive frequency bands, where each band-specific trend is modeled via spline bases whose complexity is jointly selected with the number of bands. Model complexity is controlled through a minimum description length (MDL) criterion that balances fit and parsimony, and a genetic algorithm is implemented to efficiently explore the large combinatorial space of band endpoints and spline knot configurations. Simulation studies demonstrate that the procedure accurately recovers frequency band structure and time-varying summaries, with improved performance as the number and length of replicates increase. Applied to pupil diameter time series collected during a visuospatial working memory task, the method identifies physiologically meaningful bands and reveals temporal patterns of spectral power that discriminate ADHD from control subjects, yielding high classification performance when used as input to downstream logistic and tree-based classifiers.

Keywords: Nonstationary time series, Spectral analysis, Biomedical signals.

Data Science and generative-AI

H. Brydon[a] and R. Luus[a]

[a] *University of the Western Cape, South Africa*

There are multiple definitions of Data Science but academically, we tend to view this definition as cross-disciplinary in that it is the culmination of knowledge and skills across the disciplines of Mathematics, Statistics and Computer Science. The arrival and development of generative-AI technologies within the last few years, challenges us to rethink some of the core pedagogies within this definition of Data Science.

This is not to say that the respective disciplines require change within this definition of Data Science but perhaps they do need a rethink in terms of how generative-AI would complement their usage in practise. The research presented here looks at an application and usage of generative-AI in the development and preparation of data for Statistical modelling.

The research presented will showcase how generative-AI can be used to efficiently develop code scripts for simulating data with a wide variety of characteristics. This development process further identified potential areas where critical thinking becomes even more pertinent than before and comments to this end will be included in the discussion.

Keywords: Data science, generative-AI, data simulation, critical thinking.

Interactive calibrated-axes biplots in R with the bipl5 package

R. Buys[a], S. Lubbe[a] and M. L. Steyn[a]

[a] *Stellenbosch University*

The bipl5 package provides an interactive framework in R for constructing calibrated-axes biplots as reactive HTML widgets using Plotly and custom JavaScript. Its scope includes principal component analysis, canonical variate analysis, principal coordinates analysis, and regression biplots, combining dynamic features and embedded fit measures in a single exploratory display. Designed both as standalone software for constructing biplots and as a plotting wrapper for biplotEZ objects, bipl5 offers a flexible environment for interactive multivariate visualisation.

To address the visual clutter that occurs when calibrated axes and observations compete for the same central plotting region, bipl5 implements an algorithm that translates the axes away from the plot centre while preserving their calibration. The degree of translation can be adjusted interactively on the plot. To help users assess and manage approximation quality, the package provides predictivity-based fit measures grounded in the orthogonality framework of [1], as well as an axis-scoring procedure that evaluates each axis according to its prediction error, guiding the selection of the most informative axes for display. These diagnostics are available directly within the plot, rendering a dashboard to navigate the multivariate display.

The presentation will illustrate the functionality of bipl5 and show how classical biplot methodology can be extended into a reactive environment that combines readability, interpretability, and formal fit diagnostics.

Keywords: Biplots, Interactive visualisation, Multivariate analysis.

References

[1] S. Gardner-Lubbe, N. J. le Roux, and J. C. Gower (2008). Measures of fit in principal component and canonical variate analysis. *Journal of Applied Statistics*, **35**(9), 947–965.

A Supervised Distance Metric for K-Nearest Neighbors with Mixed-Type Data

C. Cavicchia[a], M. van de Velden[a], A. Iodice D'Enza[b] and A. Markos[c]

[a]*Erasmus University Rotterdam,* [b]*University of Naples Federico II,* [c]*Democritus University of Thrace*

K-Nearest Neighbors (KNN) is a widely used nonparametric method for classification and regression. It predicts the response of a new observation from the responses of the K closest training observations in the feature space. Its effectiveness depends critically on how distances between observations are defined [1]. This is especially important in heterogeneous datasets containing both numerical and categorical variables. Standard distances are designed for numerical data and are therefore not well suited to mixed-type settings. A classical alternative is to combine variable-specific contributions into a single measure. However, such distances typically do not incorporate information from the response variable, and they may also favor one variable type over another by construction [2]. In this paper, we introduce a supervised distance for heterogeneous data that combines variable-type-specific dissimilarities with response-based weighting. The proposed distance treats numerical and categorical variables differently while assigning greater importance to predictors that are more strongly associated with the response. As a result, observations with similar responses are encouraged to be closer, whereas observations with different responses tend to be farther apart. By incorporating response information directly into the distance construction, the proposed approach improves the discriminative power of KNN for mixed-data settings. The resulting framework is simple, interpretable, and suitable for applications in which heterogeneous predictors are common.

Keywords: K-Nearest Neighbors, mixed-type data, distance metric

References

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

[2] M. van de Velden, A. Iodice D'Enza, A. Markos, and C. Cavicchia (2024). Unbiased mixed variables distance. *arXiv preprint arXiv:2411.00429*.

Unifying Weights and Loadings in Sparse Component Analysis via Equality and Cardinality Constraints

Hetvi Chaniyara[a] and Katrijn Van Deun[b]

[a] *Technical University of Berlin,* [b] *Tilburg University*

As high-dimensional, low-sample-size (HDLSS) datasets become increasingly common, there is a growing reliance on sparse principal component analysis (SPCA) and sparse factor analysis (SFA) for dimension reduction and the exploration of underlying latent structures. Their primary objective is to obtain interpretable low-dimensional representations through sparse components. Existing approaches differ in how sparsity is imposed. Weight-based SPCA methods yield explicit component scores but often suffer from instability in variable selection, as markedly different sparse weight vectors can produce nearly identical scores, a problem aggravated in HDLSS settings. Loading-based SFA methods instead promote simple structure and typically yield stable loading patterns, yet lack uniquely defined component scores due to factor score indeterminacy. Hybrid formulations include both weights and loadings but retain ambiguity regarding which parameters should guide interpretation.

We propose a sparse component method that resolves this tension by imposing equality between weights and loadings within a single constrained optimization framework. This constraint eliminates interpretationpal ambiguity, ensures uniquely defined component scores, and improves stability of variable selection. To further promote simple structure, a cardinality constraint is imposed instead of shrinkage-based regularization, motivated by its lower estimation bias and favorable support recovery properties in sparse PCA settings.

Estimation is performed using an alternating optimization scheme combining the Alternating Direction Method of Multipliers (ADMM) with Majorization–Minimization, enabling efficient computation in high-dimensional settings. Simulation studies and an empirical application demonstrate that the proposed approach yields stable and interpretable sparse components while simultaneously providing well-defined component scores. The method is com-

pared to well-known sparse PCA methods.

Keywords: Sparse PCA, Constrained optimization.

A modified two one-sided test for reference-scaled average bioequivalence

C. Chiang[a] and C.F. Hisao[b]

[a] *Tamkang University, New Taipei, Taiwan,* [b] *National Health Research Institutes, Taiwan*

The US FDA has released the first revision of the draft guidance, "Statistical Approaches to Establishing Bioequivalence," in late 2022. This revision recommends reference-scaled average bioequivalence (RSABE) for in-vitro permeation tests and assessments of narrow therapeutic index drugs. The FDA suggests that the corresponding test be based on an upper confidence bound using the Howe method for a linear combination of squared mean difference and variance(s). Since the statistical characteristics for this confidence bound are difficult to derive, the sample size determination often relies on simulations. In this study, we proposed a modified two one-sided test (MTSOT) as an alternative, accompanied by a formal method for sample size determination. Simulation results demonstrate that the proposed MTOST effectively controls the Type I error rate and provides sufficient power. Furthermore, it is more powerful than the Howe method under various scenarios.

Keywords: Reference-scaled average bioequivalence; Sample size determination; Two one-sided test

Scalable Estimation of Large Generalized Additive Models

M. Zimmermann[a], C. Collarin[b], S. N. Wood[b], F. Ziel[a]

[a]*Data Science in Energy and Environment, University of Duisburg-Essen, Germany,* [b]*School of Mathematics, University of Edinburgh, United Kingdom*

Generalized Additive Models (GAM) relate a univariate exponential family response y_i to a linear predictor $g(\mu_i) = \eta_i = \mathbf{A}_i\gamma + \sum_j f_j(x_{ij})$, where smooths f_js are represented in terms of basis functions. As a consequence, the model can be expressed as $\eta = \mathbf{X}\beta$. The p regression coefficients, $\beta = (\beta_1, \ldots, \beta_p)$, are estimated by penalized maximum likelihood: $\hat{\beta} = argmax_\beta[\ell(\beta) - (2\phi)^{-1}\sum_j \lambda_j \beta^\top \mathbf{S}_j \beta]$, where each λ_j controls the smoothness measured by $\mathbf{S}_j$. The λ_js can be estimated by Laplace approximate marginal likelihood (LAML) maximization. However, LAML requires forming and decomposing the Hessian matrix of the penalized log-likelihood. This generally precludes reducing the computational cost below $O(np^2)$, thereby making model estimation unfeasible for high-dimensional predictors.

Here we propose a way around this bottleneck. We combine the generalized Fellner–Schal smoothing parameter update with stochastic trace estimation (e.g., Hutch++, [1]) and preconditioned conjugate gradients, thus avoiding the formation or Cholesky factorisation of the GAM penalized Hessian. The resulting procedure relies only on matrix–vector products, enables low memory-bandwidth parallelization, exploits model term sparsity with minimal fill-in, and achieves an $O(np)$ computational cost. The performance of the approach is demonstrated on the NMMAPS respiratory mortality data with over one million observations and more than 20,000 coefficients, fitted in just over half an hour.

Keywords: Generalized Additive Models; Preconditioned Conjugate Gradient; Stochastic Trace Estimation

References

[1] R. A. Meyer, C. Musco, C. Musco, D. P. and Woodruff (2021). Hutch++: Optimal stochastic trace estimation. In: *Proc SIAM Symp Simplicity*

Algorithms, Jan, 142 – 155.

Graphical LASSO for Estimating Associations in Elliptically Symmetric Distributions: An EM-Based Approach

A. B. A. Dawod[a] and Gy. Terdik[b]

[a] *Doctoral School of Informatics, University of Debrecen,*
[b] *Department of Information Technology, University of Debrecen*

Modern data science applications frequently involve **high-dimensional data** that violates Gaussian assumptions. By encoding conditional dependencies within graphical models, concentration matrices enable interpretable statistical learning and exploratory data analysis.

This study develops a graphical modeling framework for non-Gaussian data based on elliptically symmetric and linear-predictor models. Within these families, zero partial correlations imply zero conditional correlations, ensuring that graph edges retain their interpretation as conditionally uncorrelated relationships. We therefore propose a modified GLASSO approach for estimating sparse concentration matrices for the generalized hyperbolic and power exponential families. By adapting an EM-type algorithm, the method enables efficient and scalable estimation in high-dimensional settings [1]. The proposed method is evaluated using simulation studies and real-data applications to assess its ability to recover underlying graphical structures under departures from normality. The results indicate that the modified framework provides improved robustness while maintaining accurate structure recovery in heavy-tailed and non-Gaussian settings [2].

Overall, this work contributes to the development of robust graphical modeling tools for modern data science, bridging statistical methodology and computational implementation.

Keywords: GLASSO, Elliptical distributions, EM algorithm.

References

[1] M. Finegold, M. Drton (2014). Robust graphical modeling with t-distributions. doi: 10.48550/ARXIV.1408.2033.

[2] R. Riccobello, G. Bonaccolto, P.J. Kremer, P. Sobczyk, M. Bogdan and

S. Paterlini, (2025). Sparse graphical modelling for global minimum variance portfolio. Computational Management Science 22(2):8.

Sparse Feature Group K-Means

A. W. Diallo[a], M. Ouattara[b], N. Niang[c] and M. Bousso[d]
[a],[d] *Université Iba Der Thiam, Thies, Senegal*
[b] *Université Polytechnique de San Pedro, San Pedro, Côte d'Ivoire*
[c] *CEDRIC – CNAM, Paris, France*

We address the problem of clustering high-dimensional multiblock data, where features are organized into homogeneous blocks representing different views of the data. Traditional subspace clustering methods like Entropy Weighting K-Means (EWKM) [1] and Feature Group K-Means (FGKM) [2] assign continuous weights to features or feature blocks, requiring tedious post-hoc analysis to identify cluster-relevant elements based on weight magnitudes. Sparse clustering methods such as Sparse K-Means [3] and Sparse Subspace K-Means [4] offer automatic feature selection but do not exploit the block structure inherent in multiblock data.

To address this limitation, we propose SFGKM (Sparse Feature Group K-Means), which extends Sparse Subspace K-Means (SSKM) to incorporate multiblock data structure. SFGKM employs a unified optimization criterion with dual-level lasso-type penalties that simultaneously performs observation clustering, cluster-specific individual feature selection, and cluster-specific feature block selection. The method automatically sets irrelevant feature and block weights to zero while assigning large weights to cluster-characterizing elements, eliminating manual weight analysis.

The optimization follows an alternating maximization scheme that iteratively updates cluster assignments, feature weights within blocks, and block weights using soft-thresholding operators until convergence. Experimental validation is conducted on synthetic datasets with varying complexity and noise patterns, as well as on real-world multiblock datasets covering diverse domains and block structures. Results show competitive performance on synthetic data, where SFGKM correctly identifies cluster-specific relevant blocks and automatically zeros out noise features. On real-world data, SFGKM achieves superior clustering performance compared to existing methods, demonstrating that explicitly modeling block structure is essential when views must be treated as coherent units rather than collections of independent features.

Keywords: Sparse Clustering, Soft Subspace Clustering, Feature Selection, Multi-block data.

References

[1] L. Jing, M. Ng, J. Huang (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, 1026–1041.

[2] X. Chen, Y. Ye, X. Xu, J. Z. Huang (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 434–446.

[3] D. M. Witten, R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 713–726.

[4] A. W. Diallo, N. Niang, M. Ouattara (2021). Sparse subspace k-means. In *ICDMW 2021*, 678–685.

Humans, Machines, and the Space Between: Engineering Effective Human-AI Partnerships

Menna El-Assady[a]

[a] *ETH Zurich*

Intelligence augmentation through mixed-initiative systems promises to combine AI's computational efficiency with human contextual expertise. In this talk, I will explore a framework for bridging human and artificial intelligence through three core pillars: interpretability, feedback, and augmentation. Central to this vision are co-adaptive visual interfaces that facilitate seamless collaboration. I will first examine how interpretable design allows users to diagnose and understand complex models, building the transparency necessary for trust. Building on this, I will discuss methods for integrating diverse human feedback directly into the learning loop, enabling models to refine their behavior in real-time. Finally, I will demonstrate how these combined elements lead to true intelligence augmentation, presenting workflow designs for computational linguistics that empower humans to solve problems more effectively than either party could alone. The talk will conclude with insights into current challenges and the promising research directions that lie ahead in engineering these symbiotic partnerships.

Globally Aligned Principal Component Analysis for Multi-Group Data

H. Fathi[a], M.A. Cremona[a] and F. Severino[a]

[a] *Université Laval, Canada*

We propose a novel principal component analysis (PCA) for multi-group datasets, where multiple numerical variables are measured across different groups. Our method combines group-specific principal components with global ones through an explicit alignment mechanism based on regularized optimization. We introduce the notion of a globally aligned covariance matrix that incorporates weighted contributions from global principal directions. In this way, we balance the preservation of within-group variance with global coherence. The alignment strength is controlled by regularization parameters that can be tuned to achieve the desired trade-off. Through a comprehensive simulation study, we demonstrate that the aligned approach achieves a favorable compromise between capturing local variation within groups and maintaining interpretability and stability across groups. The method addresses a fundamental gap in the PCA literature. Existing approaches either ignore group structure entirely, focus exclusively on local structure (group-wise PCA), or impose restrictive assumptions of common principal components. Our approach respects the multi-group nature of data while ensuring global comparability of components. Furthermore, in an application to the 2021 Canadian Census socioeconomic data, the proposed alignment yields more comparable and stable region-specific components than pooled or purely region-wise PCA.

Keywords: Principal Component Analysis, Multi-group data, Dimension reduction. (Use at most 3 keywords)

Bayesian Variable Selection in Generalized Linear Models

L. Filippozzi[a,b], I. Urteaga[c,d] and C. Agostinelli[a]

[a] *University of Trento,* [b] *Fondazione Bruno Kessler (FBK),* [c] *Basque Center for Applied Mathematics (BCAM),* [d] *Ikerbasque — Basque Foundation for Science*

Variable or covariate selection is a crucial step aimed at identifying the most relevant predictors for explaining the response variable.

In this work, we propose `BayesVS-GLM`, a novel Bayesian covariate selection method for Generalized Linear Models (GLMs). `BayesVS-GLM` is based on a fully conjugate Bayesian hierarchical model that comes with theoretical posterior consistency guarantees. Specifically, we extend the standard GLM framework by introducing a binary vector z to indicate which covariates are included in the generalized linear predictor. The regression coefficients β are modeled conditionally on z, using conjugate priors for GLMs [4].

Although related method exists ([1], [2], [3]), our method is, to the best of our knowledge, the first that provides a unified framework in which: (i) the formulation of the hierarchical GLM is fully conjugate; (ii) the GLM likelihood is explicitly dependent on indicator variables z, enabling a regressor selection based uniquely on observed data; and (iii) the posterior asymptotical accuracy of z and posterior consistency of the regression coefficients β are guaranteed. For posterior inference, we present an efficient Gibbs sampling algorithm, based on a fully conjugate Bayesian hierarchical model.

The `BayesVS-GLM` formulation is applicable to any distribution within the exponential family, and unifies a range of existing Bayesian variable selection perspectives within a single coherent hierarchical framework.

Keywords: Bayesian Variable Selection, Generalized Linear Models, Posterior model selection.

References

[1] L. Kuo, and B. Mallick (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, **60**(1), 65–81.

[2] P. Dellaportas, J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. In *Statistics and computing*, **12**(1), 27–36.

[3] N. N. Narisetty, S. Juan, and H. Xuming (2019). Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection. In *Journal of the American Statistical Association*, **114**(527), 1205–1217.

[4] M. Chen, and J. G. Ibrahim (2003). Conjugate priors for generalized linear models. In *Statistica Sinica*, **13**(2), 461–476.

Robust data cleaning for tensor-valued observations, with application to GC×GC-MS data

P. Filzmoser[a], L. Micheler[a,b], N. Lim[c], and E. Rosenberg[c]

[a]*Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria,* [b]*AC2T research GmbH, Austria,* [c]*Institute of Chemical Technologies and Analytics, TU Wien, Austria*

Nowadays, many measurement devices lead to matrix-valued (e.g. images) or even tensor-valued data. An example for the latter are data from two-dimensional gas chromatography (GC) coupled with mass spectrometry (MS), so-called GC×GC-MS data, where for each mass number an image is observed with mass intensity values along a first and second retention time. Some of these hundreds of images contain data artifacts caused by the measurement process. For reliable data analysis it is important to correct the data first, which means that images with artifacts need to be identified.

This problem boils down to identifying outlying matrices in tensor data. For that purpose we assume that the slices of the tensor follow a matrix normal distribution, and we estimate the parameters by the Matrix Minimum Covariance Determinant (MMCD) estimator [1].

After removing/correcting outlying slices for each sample, we can continue working with the cleaned tensor-valued observations. In this application the observations are measurements of fuel samples originating from different fuel types. In order to characterize chemical differences among the types we use tensor-PCA [2], where the loadings indicate mass numbers that allow to distinguish the fuel types. This form of automated procedure characterizing chemical differences for GC×GC-MS data is novel.

Keywords: Robustness, Tensor, PCA.

References

[1] M. Mayrhofer, U. Radojičić, and P. Filzmoser (2025). Robust covariance estimation and explainable outlier detection for matrix-valued data. *Technometrics*, textbf67(3) (2025), 516–530.

[2] J. Virta, S. Taskinen, and K. Nordhausen (2016). Applying fully tensorial ICA to fMRI data. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1-6.

Depth functions for tree-indexed data

G. Francisci[a] and A. N. Vidyashankar[b]

[a] *University of Trento,* [b] *George Mason University*

Depth functions quantify the degree of centrality of a point relative to a multivariate data set, with higher depth indicating greater centrality and lower depth indicating peripheral or outlying positions. They are important tools in non-parametric and robust statistics and have been used for classification and clustering. Several extensions exist for functional and metric space valued data. In this work, we introduce depth functions for tree-indexed data. Our analysis is based on the intensity measure of the point processes generating the data. When the point processes have independent and identically distributed components, the depth function reduces to the depth of a single component. We investigate the statistical properties and asymptotic behavior of the depth function. This enables the study of medians and quantiles of tree-indexed data. Finally, we apply these results to the classification of tree-indexed data using an analog of the depth-versus-depth (DD) classifier.

Keywords: Classification, Depth functions, Tree-indexed data

References

[1] G. Francisci and A. N. Vidyashankar (2024), Functional limit laws for the intensity measure of point processes and applications, *arXiv preprint arXiv:2402.05087*.

Mathematical Optimization for Scalable Scientific Visualization

S. Frey

University of Groningen, The Netherlands

Scientific data is growing rapidly in memory footprint, resolution, parameter space, and ensemble size. This offers new opportunities for insight, but also introduces visualization challenges regarding scalability, interactivity, and human perception. Mathematical optimization provides a powerful framework for addressing these issues by framing visualization tasks as problems that can be solved efficiently and systematically.

In this talk, we explore how optimization is used in scientific visualization. We discuss the kinds of problems tackled in this context, such as data reduction, layout computation, and tuning of mapping and rendering parameters—and examine the objectives and evaluation criteria involved, from technical measures like runtime to quality metrics. We also consider the requirements and trade-offs across scenarios, from real-time interaction to offline processing, and how optimization results are communicated to users.

We complement our overview with concrete examples from the visualization community and own prior work on optimization-driven visualization. Looking ahead, we highlight emerging challenges and trends, such as tighter integration with machine learning, that will enable visualization systems to handle increasingly complex and large datasets more efficiently.

Keywords: Scientific Visualization, Optimization, Large Data.

Bias–Variance Propagation from Precision Matrix Estimation to Minimum-Variance Portfolios

A. Fulci[a]

[a] *Department of Economics and Management, University of Trento*

This paper studies how estimation error in covariance and precision matrices propagates to the out-of-sample risk of the global minimum-variance portfolio. Using a Delta-method expansion, we derive a second-order bias–variance decomposition of the portfolio's out-of-sample variance and show how the variance and bias of covariance and precision estimators map into excess portfolio risk. Within this framework we obtain an oracle intensity for linear covariance shrinkage that is tailored to minimizing out-of-sample variance and generally differs from the intensity that is optimal under standard mean-squared-error criteria for covariance estimation. Then, we propose a simple variance-targeted rule that selects the shrinkage intensity by minimizing a proxy of the portfolio's out-of-sample variance on an internal validation block and rescales it to account for the larger outer estimation window.

A Monte Carlo study based on factor-structured data-generating processes, together with an empirical study on three large equity universes (S&P 100, S&P 500 and STOXX Europe 600), compares this procedure with the Ledoit–Wolf and Schäfer–Strimmer covariance estimators, and shows that the variance-targeted rule uses higher shrinkage intensities and typically achieves lower out-of-sample volatility and turnover.

Keywords: global minimum-variance portfolio, covariance shrinkage, precision matrix estimation, bias–variance trade-off, cross-validation, out-of-sample risk, portfolio turnover

Multivariate Wrapped Normal estimation with missing values

L. Greco[a] and C. Agostinelli[b]

[a] *University Giustino Fortunato Benevento, Italy,* [b] *Department of Mathematics, university of Trento, Italy*

This contribution addresses maximum likelihood estimation estimation and model-based imputation for multivariate circular data lying on a p-dimensional torus in the presence of missing values, when the missing data mechanism is ignorable [2]. Actually, the periodic nature of the sample space invalidates conventional imputation techniques designed on the Euclidean space. The multivariate Wrapped Normal distribution (WN) is a flexible and computationally tractable model for torus data, with pdf $\phi_p^{\circ}(y;\mu,\Sigma) = \sum_{j\in\mathbb{Z}^p} \phi_p(y + 2\pi j;\mu,\Sigma)$, where $\phi_p(\cdot;\mu,\Sigma)$ denotes the p-variate normal density function, $y \in [0,2\pi)^p$, $\mu \in [0,2\pi)^p$ is the mean direction, $\Sigma \in \mathbb{R}^{p\times p}$ is a positive definite scatter matrix. In practice, the infinite sum is truncated to a finite set $\mathcal{C}_J = \{-J, -J+1, \ldots, J\}^p$ for a sufficiently large J, as the terms decay rapidly for concentrated distributions [3, 1]. The methodology leverages the conditional properties of the normal distribution on the unwrapped space, embedding the imputation of missing torus data into an Expectation-Maximization algorithm that treats both the wrapping coefficients and the missing entries as latent variables.

Keywords: EM algorithm, MAR, MCAR.

References

[1] Luca Greco, Pier Luigi Novi Inverardi, and Claudio Agostinelli. Finite mixtures of multivariate wrapped normal distributions for model-based clustering of p-torus data. *Journal of Computational and Graphical Statistics*, 32(3):1215–1228, 2023.

[2] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, 3rd edition, 2019.

[3] Ahmad Nodehi, Mohsen Golalizadeh, Mehdi Maadooliat, and Claudio Agostinelli. Estimation of parameters in multivariate wrapped models

for data on a p-torus. *Computational Statistics*, 36(1):193–215, 2021. doi: 10.1007/s00180-020-01006-x.

Interpretable Kernels

P. J. F. Groenen[a] and M. Greenacre[b]

[a] *Econometric Institute, Erasmus University Rotterdam,* [b] *Universitat Pompeu Fabra*

The use of kernels for nonlinear prediction is widespread in machine learning. They have been popularized in support vector machines and used in kernel ridge regression, amongst others. Kernel methods share three aspects. First, instead of the original matrix of predictor variables or features, each observation is mapped into an enlarged feature space. Second, a ridge penalty term is used to shrink the coefficients on the features in the enlarged feature space. Third, the solution is not obtained in this enlarged feature space, but through solving a dual problem in the observation space. A major drawback in the present use of kernels is that the interpretation in terms of the original features is lost. In this paper, we argue that in the case of a wide matrix of features, where there are more features than observations, the kernel solution can be re-expressed in terms of a linear combination of the original matrix of features and a ridge penalty that involves a special metric. Consequently, the exact same predicted values can be obtained as a weighted linear combination of the features in the usual manner and thus can be interpreted. In the case where the number of features is less than the number of observations, we discuss a least-squares approximation of the kernel matrix that still allows the interpretation in terms of a linear combination. It is shown that these results hold for any function of a linear combination that minimizes the coefficients and has a ridge penalty on these coefficients, such as in kernel logistic regression and kernel Poisson regression. This work makes a contribution to interpretable artificial intelligence.

Keywords: Interpretable artificial intelligence, Kernel regression, Ridge regression.

References

[1] Groenen, P.J.F., and Greenacre, M. (2025). Interpretable Kernels. *arXiv preprint*, arXiv:2508.15932.

Advancing Clinical Research with Large-Scale EHR Data and Machine Learning

Yi Guo, PhD, FAMIA[a]

[a] *Department of Health Outcomes and Biomedical Informatics, University of Florida*

The widespread adoption of electronic health record (EHR) systems has created unprecedented opportunities for data-driven clinical research by enabling access to large-scale, longitudinal real-world data (RWD). National initiatives such as PCORnet have demonstrated how harmonized EHR and claims data can support population-scale analytics across diverse health systems. As one of the eight PCORnet networks, the OneFlorida+ Clinical Research Network maintains a centralized repository of linked, longitudinal RWD for over 20 million patients across multiple U.S. states. This talk will introduce the OneFlorida+ data infrastructure from a data science perspective and discuss how large-scale clinical data can be integrated with machine learning and artificial intelligence methods for tasks such as phenotyping, risk prediction, and outcome modeling. Particular emphasis will be placed on methodological considerations, including data heterogeneity, temporal structure, bias, and causal interpretation, that are critical for translating advanced analytics into robust and trustworthy clinical insights.

Keywords: Real-World Data, Machine Learning in Healthcare, Causal Inference.

Research on Data Assets Pricing Based on LLM Agent

Haibo Han

Lanzhou University of Finance and Economics

Conventional supply-side pricing models fail to capture the dynamic value of data assets in complex markets. This dissertation proposes a multi-agent AI framework where specialized autonomous agents collaborate to determine data asset prices through tool-augmented reasoning and collective deliberation.

The system comprises Value Assessment, Market Analysis, and Risk Evaluation agents coordinated by an Orchestrator. Each agent autonomously retrieves information through external tools and APIs, while a multi-agent debate protocol resolves valuation conflicts. A three-tier memory system enables continuous learning from historical cases.

Empirical results demonstrate superior performance: R^2 of 0.8473 (vs. 0.8154 for single-agent and 0.6423 for traditional ML), with multi-agent collaboration contributing 18.7% and tool augmentation 24.3% to overall accuracy. The debate mechanism reduces valuation variance by 34.2%, while human-in-the-loop verification achieves 96.8% expert concordance.

This research establishes a novel paradigm for autonomous economic decision-making, integrating multi-agent collaboration and human-AI symbiosis for robust data asset pricing.

Keywords: Data Assets; Pricing Model; Deep Learning; Large Language Model; Demand-side Pricing

On the Role of Interaction in Visual Data Science

Helwig Hauser
University of Bergen, Norway

Interactive visual data exploration – also referred to as exploratory data analysis, EDA – is common in data science, not only as part of hypothesis generation, but also when prototyping appropriate data analysis workflows. Interactive data visualization is a key ingredient and a solid foundation of useful methods – both regarding appropriate visual representations as well as user interaction techniques – has been established and evaluated in the scientific field of visualization and visual analytics, including coordinated multiple views, linking and brushing, focus+context visualization, etc.

In this contribution, we take a closer look at the role of interaction in visual data science. We examine potential benefits – for example, how interactive visualization becomes part of externalization, enabling and supporting complementary levels of cognition during data exploration – as well as possible downsides (human-in-the-loop processes are potentially costly, requiring the user's time, for example). We discuss, how interactive visualization can be understood as a form of dialogue between the analyst and her/his data and we analyze, which requirements – primarily in the temporal sense – are faced when establishing an effective and efficient interaction.

Besides an overview of selected examples from published work in visualization research, we also considered a few concrete examples from own prior work on interactive visual data exploration and analysis, including a demonstration of how deep learning can help to optimize user interaction in visualization.

Keywords: visual data science, interactive visualization, interaction.

On stochastic structure of thinning-based non-negative vector-valued ARMA processes

M. Ispány

Faculty of Informatics, University of Debrecen, Hungary

Count time series models are widely used in practice, see the survey [1]. In the talk, we provide a unified, distribution-free discussion of the following thinning-based non-negative integer vector-valued ARMA model

$$Y_k = \sum_{i=0}^{p} A_i^k \circ Y_{k-i} + \sum_{j=0}^{q} B_j^k \circ \varepsilon_{k-j}, \quad k \in \mathbb{Z}, \tag{1}$$

where $\{A_i^k\circ\}$, $i = 0, 1, \ldots, p$, and $\{B_j^k\circ\}$, $j = 0, 1, \ldots, q$, are independent identically distributed matricial thinning operators of dimension $d \times d$, respectively, and the input process $\{\varepsilon_k\}$ is a sequence of independent identically distributed $\mathbb{N}_0^d$-valued random vectors. The matricial thinning operators and the random vectors are mutually independent. The solution $\{Y_k\}$ of (1) is a generalized branching process called INVARMA(p, q) process, see [2] for an example.

We give a simple moment assumption and a spectral criterion involving the model parameters that ensures the existence and uniqueness of a solution to (1). We present a graphical model for describing the stochastic structure of the process and the dependencies among offspring generations and the moving average part of the model as a one-sided infinite skew comb.

Joint work with P. Bondon (CentraleSupélec, France), V. A. Reisen (UFES, Brazil).

Keywords: ARMA model, count time series, graphical model.

References

[1] R. A. Davis, K. Fokianos, S. H. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, N. Ravishanker (2021). Count Time Series: A Methodological Review. *Journal of the American Statistical Association*, **116**(535), 1533–1547.

[2] M. Ispány, P. Bondon, V. A. Reisen, P. R. P. Filho (2024). Existence of a periodic and seasonal INAR process. *Journal of Time Series Analysis*, **45**(6), 980–1005.

Interactive Visual Parameter Exploration in Functional Data Analysis

Truls H. Jakobsen [a], Ana-Maria Urdea [b], Rainer Splechtna [c], Krešimir Matković [c], Peter Filzmoser [b], Helwig Hauser [a]

[a] *Univ. of Bergen, Norway*; [b] *TU Wien, Austria*; [c] *VRVis Research Center, Austria*

Functional data analysis (FDA) treats observations as functions rather than as isolated measurements [1], and a central challenge in any FDA workflow is the configuration of the function representation itself. Analysts must choose a set of basis functions, determine the appropriate parameterization of that basis (number of basis functions, order, knot placement, and so on), and select an adequate smoothing level before they can inspect the consequences of their choices in any detail. In script-driven environments, revisiting these decisions requires rewriting and rerunning code, making it difficult to develop intuition about model sensitivity or to compare alternatives efficiently.

We present a prototype that addresses this by connecting R directly to an interactive visualization application. We integrate ComVis, a coordinated multiple views framework for prototyping visualization technology [1], with the `fda` package [1] as a statistical backend.

The result is a semi-guided workflow in which parameter choices are exposed as interactive controls, and linked views provide immediate visual feedback on how each choice affects the fitted functions and their derivatives. Analysts can brush subsets of interest and observe modelling consequences without leaving the visual environment. Immediate visual feedback guides the analyst toward an appropriate FDA pipeline configuration, and the ability to display multiple parameter configurations simultaneously provides a direct means of comparison and a clearer understanding of parameter influence.

Keywords: Functional data analysis, interactive visualization, visual analytics

References

[1] Krešimir Matković, Wolfgang Freiler, Denis Gračanin, and Helwig Hauser. Comvis: A coordinated multiple views system for prototyping

new visualization technology. In *12th Int'l Conf. on Information Visualisation*, pages 215–220, 2008. doi: 10.1109/IV.2008.87.

[1] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2005. doi: 10.1007/b98888.

Beyond Binary Ties: Modeling the Evolution of Relational States

R. Juozaitienė

Vytautas Magnus university

Researchers across diverse domains increasingly consider various phenomena from a dynamic network perspective, viewing systems as composed of nodes connected by ties that evolve over time. Most existing modelling approaches, however, rely on binary representations in which relationships are either present or absent. Such representations overlook the heterogeneity and varying intensities that characterize many real-world ties. In practice, relationships often unfold through multiple substantively meaningful states, such as acquaintanceship, friendship, or collaboration, that reflect relatively stable relational conditions rather than isolated interactions. Reducing these states to a binary structure loses important information about how ties emerge, intensify, and dissolve.

To address this limitation, we introduce a continuous-time framework for modelling dynamic networks in which each dyad occupies one of several distinct relational states and evolves through transitions between them. Building on relational event models, we treat state transitions as probabilistic events and model the waiting times between them using hazard-based approaches. Transition rates are allowed to depend on endogenous network structures—such as reciprocity and triadic closure—as well as exogenous covariates. Extending relational event models to multi-state settings requires novel definitions of endogenous network effects and estimation strategies capable of capturing the added structural complexity. The proposed framework provides a flexible approach for analyzing the evolution of relational states while preserving the substantive richness of network dynamics.

Keywords: multi-state ties, relational states, dynamic networks

Centralised validation and AI report generation: Accelerating statistical dissemination at an NSO

A. Kanyesigye, I. Atwiine, F. Kayondo, and L. Mugula

Uganda Bureau of Statistics (UBOS), Kampala, Uganda

A National Statistical Office holds vast quantities of data, yet producing a single policy brief can take days. The bottleneck is not computation: it is that the underlying data lives in disconnected, incompatible systems. Poverty survey results arrive as Excel spreadsheets, census figures as PDF reports, price indices from administrative databases, sector statistics as SDMX feeds. Each has a different schema, update cadence, and quality profile. Before any analyst can write a paragraph, they must locate, extract, reconcile, and manually validate figures from several of these sources simultaneously. This paper presents a pipeline developed at the Uganda Bureau of Statistics (UBOS) that eliminates this bottleneck by combining automated multi-source validation with AI-driven report generation.

The pipeline ingests data from all source systems into a single centralised analytical layer through format-specific parsers and automated type detection. Before any data is admitted, it passes through a validation toolkit that runs structural checks (schema conformity, type enforcement, metadata completeness), statistical checks (range validation, confidence interval reconciliation, 3-sigma anomaly detection per indicator and disaggregation level), and a cleaning stage that logs every transformation with a full audit trail. A key principle is that no value is silently altered: rows requiring human judgment are tagged and quarantined rather than modified. A weighted quality score across structural, statistical, and conformity dimensions gates data into the centralised layer, which is maintained as a single trusted source of truth aligned to SDMX metadata standards.

Report generation operates directly against this centralised layer. Given a target document type — policy brief, statistical bulletin, or press release — a large language model retrieves the required indicators, generates data-grounded visualisations (bar charts, trend lines, maps via Plotly/Matplotlib), and produces structured prose following a configurable schema. Because the

AI operates on already-validated, centralised data, there is no per-document extraction or re-validation step. What previously required days of manual compilation from disparate systems is reduced to minutes, and the same centralised layer can serve multiple document types concurrently.

The system is demonstrated on Uganda National Household Survey poverty indica- tors and 2024 Census preliminary results. We report end-to-end generation time versus manual authoring, quality score distributions across ingested sources, and domain-expert ratings of AI-generated policy briefs on accuracy, completeness, and usability. The results show that centralised, validated data is not merely a quality improvement — it is the architectural prerequisite that makes fast, reliable AI dissemination possible.

Keywords: Statistical data integration, Report generation, Data validation.

References

[1] J. Reis, and M. Housley (2022). *Fundamentals of Data Engineering*. O'Reilly Media.

[2] M.D. Wilkinson et al. (2016). The FAIR guiding principles for scientific data management. *Scientific Data*, 3, 160018.

Deep Learning for Poverty Estimation in Côte d'Ivoire using Sentinel-2 Imagery

K. M. Koné[a], M. R. Kouamé[a], F. A. Migoné[a], D. J. Koné[a], K. A. Kouassi[a] and D. Doukouré[a]

[a] *National Statistics Agency (ANStat), Côte d'Ivoire*

This study develops a methodology for estimating poverty in Côte d'Ivoire using Sentinel-2 satellite imagery and deep learning transfer techniques. The approach follows a three-stage pipeline: pre-training a VGG16 convolutional neural network [1] on ImageNet [2], fine-tuning it to predict nighttime light intensity from daytime imagery, and extracting visual features for poverty prediction. Using data from the 2018-2019 Harmonized Household Living Conditions Survey covering 12,992 households, visual features are enriched with spectral indices including NO_2 concentration, Land Surface Temperature, and the Enhanced Normalized Difference Impervious Surface Index. The best model achieves an R^2 of 40.15%, with predictions remarkably consistent with survey data: 39.98% predicted versus 39.44% observed in 2018. Applying the 2018-trained model to 2021 imagery yields an inferred rate of 37.86%, closely matching the 37.50% survey estimate, confirming the approach's capacity to capture poverty dynamics over time. This methodology offers a cost-effective complementary tool for monitoring poverty between survey waves in Sub-Saharan Africa.

Keywords: Poverty estimation, Satellite imagery, Deep learning.

References

[1] K. Simonyan and A. Zisserman (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei (2009). ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

CougarStats: Integrating Computational Statistics and Interactive Visual Analytics

A. Krishnamurthy[a], M. Abou El Nasr [a] and J. Bennett [a]

[a] *Mount Royal University*

CougarStats (www.cougarstats.ca) is an open-source web-based platform developed in the R ecosystem that integrates statistical modelling and visualization within a unified environment for data analysis.

The platform provides access to core statistical techniques through interactive modules spanning descriptive statistics, statistical inference, regression modelling, and machine learning. Rather than treating modelling and visualization as separate steps, CougarStats supports interactive analytical workflows in which statistical estimation, diagnostic assessment, and graphical exploration occur simultaneously. This integration reduces the cognitive gap between computing a result and interpreting its significance.

A central design principle of CougarStats is computational transparency in statistical computing and visualization. Analytical workflows are implemented using established R methods while exposing step-by-step calculations and dynamically linked graphical output that facilitate interpretation of model behaviour and underlying data structure.

The platform continues to expand to support a broad range of modern data science workflows, including supervised classification methods (e.g. Linear Discriminant Analysis (LDA)) and unsupervised dimensionality reduction techniques (e.g. Principal Component Analysis (PCA)), and predictive modelling for time-series data.

Keywords: Data science, Computational Statistics, Visualization.

Copula-based models for spatially dependent cylindrical data

F. Labanca[a], A. Gottard[a] and N. Klein[b]

[a] *University of Florence,* [b] *Karlsruhe Institute of Technology*

Cylindrical data frequently arise across various scientific disciplines, including meteorology (e.g., wind direction and speed), oceanography (e.g., marine current direction and speed or wave heights), ecology (e.g., telemetry). Such data often occur as spatially correlated series of intensities and angles, thereby representing dependent bivariate response vectors of linear and circular components. To accommodate both the circular-linear dependence and spatial autocorrelation, while remaining flexible in marginal specifications, copula-based models for cylindrical data have been developed in the literature. However, existing approaches typically treat the copula parameters as constants unrelated to covariates, and regression specifications for marginal distributions are frequently restricted to linear predictors, thereby ignoring spatial correlation. In this work, we propose a structured additive conditional copula regression model for cylindrical data. The circular component is modeled using a wrapped Gaussian process [1], and the linear component follows a distributional regression model. Both components allow for the inclusion of linear covariate effects. Furthermore, by leveraging the empirical equivalence between Gaussian random fields (GRFs) and Gaussian Markov random fields [2], our approach avoids the computational burden typically associated with GRFs, while simultaneously allowing for non-stationarity in the covariance structure. Posterior estimation is performed via Markov chain Monte Carlo simulation. We evaluate the proposed model in a simulation study and subsequently in an analysis of wind directions and speed in Germany.

Keywords: Dependence structure; Gaussian (Markov) random field; Wrapped Gaussian process.

References

[1] G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics*, **6** (4), 1478 - 1498.

[2] F. Lindgren, H. Rue, and J. Lindström, (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B* **73**, 423–498.

Efficient and scalable Bayesian inference for joint modelling longitudinal and event history data

Guangquan Li

Applied Statistics and Data Science Lab, School of Engineering, Physics, and Mathematics, Northumbria University, UK

Joint models of longitudinal and event history data provide a power statistical framework for analysing complex health data from clinical studies and longitudinal studies of health. However, they are computationally expensive to fit due to the complex joint likelihood and the need to estimate many individual-level random effects. In this talk, we will introduce fastBJM, an efficient algorithm for fitting Bayesian joint models using Markov chain Monte Carlo. The algorithm updates model parameters via the Metropolis-within-Gibbs scheme, allowing us to exploit the structures of the full conditionals to derive efficient sampling. The results from our extensive simulation studies have demonstrated the estimation accuracy of fastBJM and its faster fitting compared to existing methods. The simulations also highlight its scalability in handling complex models as well as large datasets. Our application uses fastBJM to fit a suite of joint models to data extracted from the Survey of Health, Ageing and Retirement in Europe (SHARE), a large-scale multi-country longitudinal survey on health. We will discuss the results on the value of using body mass index as a tool to dynamically monitor the risks of hypertension and stroke.

Keywords: Joint modelling, Bayesian computation, Longitudinal health survey

Classification of highly colinear infrared spectral data

S. Lubbe[a,b], N. J. le Roux[a], R. J. Cornelissen[c] and H. H. Nieuwoudt[d]

[a] *MuViSU (Centre for Multi-dimensional Data Visualisation), Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa,* [b] *NITheCS (National Institute of Theoretical and Computational Sciences), Stellenbosch University, Stellenbosch, South Africa,* [c] *Namaqua Wines, Vredendal, South Africa,* [d] *South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Stellenbosch University, Stellenbosch, South Africa*

This study addresses a classification problem involving variable selection from a large set of highly correlated predictors. In spectral data, measurements at neighbouring wavenumbers are typically strongly correlated, resulting in severe multicollinearity. We propose a variable selection approach that first clusters highly correlated variables and then selects a single representative from each cluster for potential use in the classification model. The method also incorporates a constraint to prevent non-contiguous variables from being grouped within the same cluster. The approach is demonstrated using mid-infrared (MIR) spectral data from grape samples collected to assess grapevine bunch rot, an issue of economic importance to wineries. The objective is to classify new samples either as rot-affected (Yes/No) or into multiple categories reflecting the severity (%) of rot. From the smaller set of potential variables, those important for classification are selected and biplots are used to visualise the separation between classes.

Keywords: biplots, classification, multicollinearity, clustering, bunch rot

Analysing Tail Dependencies of Temperature in Agriculturally Active Regions Across South Africa.

V. N. Masingi[a], G. L. Grobler[b] and S. C. Liebenberg[a]

[a]*Focus Area for Pure and Applied Analytics, North-West University,* [b] *Unit for Data Science and Computing, North-West University*

Extreme temperature events pose significant risks to perennial crops, such as citrus, by reducing flowering at low temperatures and impairing fruit development at high temperatures. Reliable meteorological data are therefore essential for agricultural risk management. However, such datasets are often affected by missing values, measurement errors, and outliers due to technical and environmental factors, especially in agriculturally active areas. The estimation of missing extreme observations and the detection of anomalies rely on a strong dependence in the tails. However, using methods that capture dependence over the full distribution may not produce accuracy in the tails. Therefore, this study investigates the tail dependence between temperature extremes derived from agriculturally active areas in South Africa that produce citrus and other fruits. Specifically, reference temperatures measured at weather stations within these areas are compared with auxiliary observations from ERA5 and neighbouring weather stations. These comparisons are used to assess their suitability for extreme value imputation and outlier detection. This study employed two methods to analyse tail dependence, namely a naive estimation method of tail dependence coefficients and estimates from several fitted Copula models. Goodness-of-fit was assessed using likelihood-based measures computed on observations exceeding high thresholds, thereby focusing on the tails of the distribution. The study findings reveal that, for maximum temperatures, the reference observations show a considerably higher upper-tail dependence with ERA5 observations compared to the estimated lower tail dependence, suggesting an asymmetry in the dependence structure. However, this asymmetry is not as evident when comparing the estimated upper and lower tail dependencies of the reference and auxiliary weather stations. Furthermore, the results show that, where there is a strong tail dependence in the tail, the BB1 and BB7 copulas provide a good fit in the tails. This work lays the foundation for the use of

copula methods to detect outliers and impute missing meteorological data.

Keywords: Tail dependence, Copulas, extreme temperature.

Statistical modeling of high-order interactions: recent developments and future challenges

C. Matias[a,b,c]

[a] *Centre National de la Recherche Scientifique,* [b] *Sorbonne Université,* [c] *Université Paris Cité, France*

The growing interest in modeling high-order interactions (HOIs) arises from the acknowledgement that many phenomena are fundamentally more complex than what pairwise relationships alone can capture. While networks and their mathematical representation as graphs capture interactions between pairs of entities, HOI are inherently of a different nature, as they may involve the interaction of more than two elements. Taking into account HOI offers a richer and more expressive way to model complex dependencies and interactions across diverse fields, ranging from social network analysis or co-authorship relations, to ecological systems, neurosciences or even chemistry.

In this talk, we will provide an overview of recent advances in the statistical modeling of HOIs, with a particular focus on hypergraphs and the limitations inherent in graph-based representations. We will address the key issues posed by HOIs in comparison to their pairwise counterparts. Additionally, we will explore the challenges associated with node clustering and the definition of communities within this framework.

Keywords: Hypergraphs, node clustering.

When the Model Won't Invert: Interactive Visual Optimization of Complex Simulations

K. Matković[a], R. Splechtna[a] and H. Hauser[b]

[a] *VRVis Research Center, Vienna, Austria,* [b] *University of Bergen, Norway*

Many systems in science and engineering can be modeled and simulated, but not solved analytically. The model maps control parameters, etc., to model outputs through coupled, computationally expensive, non-linear equations. Such models are 'black boxes': they can be 'run' forward, but not inverted. For a desired output, there is no closed-form way to recover the parameters that produced it. This is true wherever the governing physics resist linearization and outputs are shaped by interactions that cannot be decomposed into independent, simple parts. The challenge is not merely computational. Optimization faces three compounding difficulties: the parameter space is usually high-dimensional, making dense sampling prohibitive; objectives are often conflicting, requiring trade-offs that may be user-dependent; and many critical constraints are implicit – an experienced practitioner recognizes them immediately, but they resist formal specification. We present and discuss a methodology integrating simulation, surrogate modeling, and interactive visualization: an initial ensemble of simulation runs sparsely covers the parameter space, a comparably simple surrogate model is fitted to this ensemble, enabling rapid estimation and the swift generation of new candidate solutions. These candidates are presented to the expert user through interactive, domain-familiar visualizations. The expert then explores and refines; new runs are targeted at the region of interest and the cycle continues. Every surviving candidate is checked by means of the fully-fletched simulation. This human-in-the-loop approach is necessary as implicit knowledge cannot be encoded in an objective function, but it can be exercised through a visual interface. We illustrate the approach on two use cases [1, 2], where the workflow reduced design exploration time by at least an order of magnitude.

Keywords: Visualization, surrogate modeling, interactive optimization.

References

[1] K. Matković et al., "Visual Analytics for Complex Engineering Systems: Hybrid Visual Steering of Simulation Ensembles," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, 2014.

[2] R. Splechtna et al., "Interactive Design-of-Experiments: Optimizing a Cooling

System," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 1, 2025.

Robust clustering for matrix-variate data

M. Mayrhofer[a], L. A. García Escudero[b] and A. Mayo Íscar[b]

[a] *TU Wien,* [b] *University of Valladolid*

Matrix-valued data arise naturally when a sample has more structure than a vector can carry, e.g., a grayscale image is a matrix of pixel intensities. They are often transformed into high-dimensional vectors (by stacking rows or columns), which can limit many multivariate data analysis procedures. Alternatively, they can be treated as samples from a matrix-variate distribution, which enables simultaneous modeling of row and column covariances via a Kronecker-product covariance structure.

We propose MTCLUST, a new robust clustering method for matrix-valued data, combining ideas from the recently introduced matrix minimum covariance determinant (MMCD) estimators for robust mean and covariance estimation for matrix-valued data of [1] and trimmed clustering of [2]. MTCLUST trims the most outlying samples and assigns the regular samples to one of G groups. For each group, it estimates the mean as well as the row and column covariances. Finally, it restricts the ratio between the maximum and minimum eigenvalues of the row and column covariance matrices, ensuring the problem is well-defined and simultaneously avoiding ill-conditioned covariances as well as spurious clusters.

In the single-group case, MTCLUST yields a condition-number-regularized version of MMCD. In the multi-group setting, MTCLUST has computational advantages over vectorized methods due to lower sample-size requirements, especially for initialization.

Keywords: Condition number regularization, Trimmed maximum likelihood, Spurious clusters

References

[1] Mayrhofer, M., Radojičić, U., & Filzmoser, P. (2025). Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data. *Technometrics*, 67(3), 516–530.

[2] García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A.

(2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3), 1324-1345.

On fuzzy classification indices and their variability assessment

Y. Melnykov[a], V. Melnykov[a] and M. Nai Ruscone[b]

[a] *University of Alabama,* [b] *University of Genoa*

In the area of cluster analysis, the comparison of obtained partitions is a task of primary importance. It is standard to compare the partitions based on one of available classification indices. The vast majority of such indices assume crisp membership assignments and their direct application to fuzzy procedures may yield biased results. While several attempts have been made to develop classification indices applicable in the fuzzy framework, all current approaches focus on providing a point value without taking the variability in membership assignments into consideration. We propose several novel variants of fuzzy classification indices and present an approach to derive their distributions. This allows to develop confidence intervals and hypothesis testing procedures for the introduced indices.

Keywords: Classification Index, Cluster Analysis, Fuzzy Classification.

transDA: an R Package for Transformation Discriminant Analysis

J. Li[a] and Y. Melnykov[a]

[a] *University of Alabama, Tuscaloosa AL 35487, USA*

The Discriminant Analysis, which includes Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) [1], and Mixture Discriminant Analysis (MDA) [2], is recognized for its versatile and reliable approach to classification tasks, and it has countless applications across various fields of study. However, its effectiveness is often constrained by the assumption that each group or subgroup follows a Gaussian distribution, which may not hold for real-world data. Our R package **transDA** addresses this limitation by integrating transformation into discriminant analysis, allowing for skewness within data groups or subgroups. **transDA** can handle both non-transformation methods such as LDA, QDA, and MDA and transformation methods [3, 4]. This paper provides a solid theoretical foundation about transfortamion models and detailed descriptions of the functions of **transDA**, along with illustrative example.

Keywords: Discriminant analysis, R package, Transformation models

References

[1] R. A. Fisher (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.

[2] T. Hastie and R. Tibshirani (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B*, 58, 155–176.

[3] G. E. Box and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2), 211–252.

[4] B. Manly (1976). Exponential data transformations. *Journal of the Royal Statistical Society: Series D*, 25(1), 37–42.

Fine motor performance and global cognitive function with incomplete data in the population-based CHRIS study

R. Melotti[a], M. Banerjee[a,b], A. J. Kaat[c], M. Gögele[a], F. Del Greco M.[a], P. P. Pramstaller[a], R. Gershon[c], C. Pattaro[a] and S. Wang[a,c]

[a]*Institute for Biomedicine, Eurac Research, Italy,* [b]*Department of Biostatistics, University of Michigan School of Public Health, MI, USA,* [c]*Department of Medical Social Science, Feinberg School of Medicine, Northwestern University, IL, USA*

Fine motor performance is a potential marker of cognitive impairment. Global cognitive function and its relationship with spiral drawing features were assessed under self-selection and missingness in the Cooperative Health Research in South Tyrol (CHRIS) study.

There were 8,230 out of 13,367 (61.6%) CHRIS participants without tremor related conditions and 6 replicates of the digital spiral drawing test.[1] Among these, 7,401 had full 30-item response, 216 had only partial item response, and 613 had full missing items to the Mini-Mental State Examination (MMSE). Mean drawing error from the smoothed spiral trace and median speed were averaged features of fine motor performance. A multiple imputation algorithm was developed, tested for trace convergence and applied with sampling weights to rescue missing data. Proportions of MMSE score categories [30, 29-27, 26-24, <24], as well as associations between the MMSE score and spiral features in adjusted generalized linear models were estimated either including or excluding imputed missing data.

Estimated proportions of MMSE score lowest two categories [26-24, <24] were 9.4% and 1.0%, 12.0% and 1.7%, and 12.3% and 2.2% in non-imputed, non-imputed weighted, and imputed weighted data, respectively. MMSE score had an inverse linear association with both mean drawing error and median speed in all models. However, the size of these associations was consistent between complete and imputed data for mean drawing error, not for median speed.

The proportion of low cognitive function was underestimated by self-selection and missingness in the CHRIS data. However, MMSE associations with different features of fine motor performance were not equally affected.

Keywords: CHRIS, Multiple Imputation, Selection Bias

References

[1] R. Lundin, et al. (2025). Cohort Profile: the Cooperative Health Research in South Tyrol study. *Int J Epidemiol*, **54**(3):dyaf064

A robust distance concept for Random Surfaces with application to classification

Leopold Micheler[a,b], Marcus Mayrhofer[a], Una Radojicic[a], Peter Filzmoser[a]

[a] *Institute of Statistics and Mathematical Methods in Economics, TU Wien*
[b] *AC2T research GmbH (AC^{2}T), Austria*

This contribution introduces a new distance measure and a robust method for covariance estimation of random surfaces with a separable covariance structure. The approach links separable random surfaces to the matrix-variate distribution of their basis function representations, which provides a convenient and principled framework for estimation.

Building on this connection, we develop a robust procedure based on the Matrix Minimum Covariance Determinant (MMCD) [1] estimator, combined with a truncated functional Mahalanobis semi-distance to estimate mean and covariance functions. This formulation is designed to remain stable under contamination and to provide reliable distance-based inference for functional data.

To improve interpretability, we extend the Shapley value [2] methodology to the functional setting. This allows us to decompose the proposed distance measure and thus functional outlyingness into contributions from specific spatial and temporal regions. We further introduce a novel formulation for computing these functional Shapley values while preserving their fundamental axiomatic properties.

The resulting framework integrates matrix-variate modeling, robust distance measures, and interpretable decomposition techniques into a unified approach for analyzing random surfaces. We provide theoretical justification alongside empirical results on real-world datasets, demonstrating strong performance in classification and robust outlier detection tasks.

Keywords: Functional data analysis, Robustness, Classification

References

[1] M. Mayrhofer, U. Radojičić and P. Filzmoser (2025). Robust covariance estimation and explainable outlier detection for matrix-valued data. *Technometrics*, **67**(3), 516–530.

[2] L. S. Shapley (1953). A Value for n-Person Games. *Contributions to the Theory of Games, Volume II*, 307–318.

Addressing an extreme positivity violation to distinguish the causal effects of surgery and anesthesia via separable effects

A. J. Pitts[a], L. Guo[b], C. Ing[b], and C. H. Miles[b]

[a] *Regeneron Pharmaceuticals,* [b] *Columbia University*

The U.S. Food and Drug Administration has cautioned that prenatal exposure to anesthetic drugs during the third trimester may have neurotoxic effects; however, there is limited clin- ical evidence available to substantiate this recommendation. One major scientific question of interest is whether such neurotoxic effects might be due to surgery, anesthesia, or both. Isolating the effects of these two exposures is challenging because they are observationally equivalent, thereby inducing an extreme positivity violation. To address this, we adopt the separable effects framework of Robins and Richardson (2010) to identify the effect of anesthe- sia (alone) by blocking effects through variables that are assumed to completely mediate the causal pathway from surgery to the outcome. We apply this approach to data from the na- tionwide Medicaid Analytic eXtract (MAX) from 1999 through 2013, which linked 16,778,281 deliveries to mothers enrolled in Medicaid during pregnancy. Furthermore, we assess the sensitivity of our results to violations of our key identification assumptions.

Keywords: Anesthesiology, Causal inference, Claims data, Overlap violation, Positivity violation, Separable effects

MLMC: Visualizing Multi-Label Classification

Torsten Möller

University of Vienna

Machine learning classifiers are increasingly applied to complex tasks such as audio tagging, image labeling, and text classification – many of which require multi-label classification. Traditional evaluation tools, often limited to single metrics such as accuracy, fall short of providing insight into classifier behavior across multiple labels. To address this, we designed MLMC, an interactive visualization tool for evaluating and comparing multi-label classifiers. Based on expert interviews, MLMC supports analysis at instance-, label-, and classifier-level views, offering a scalable, more interpretable alternative. I demonstrate its use across three different domains and describe its core algorithms and user interface. Two pilot studies (N=6 each) provided insight into MLMC's usability and showed improved task accuracy, consistency, and user confidence compared to confusion matrices. Results highlight MLMC's potential as a practical tool for intuitive evaluation of multi-label classifiers, with implications for a broad range of machine learning applications.

Keywords: Classification, Multi-Label, Visualization Design Study. (Use at most 3 keywords)

Contrasting disinformation: A mixture of Hawkes Processes for the identification of coordinated inauthentic behaviour

E. Muratore[a,b], R. Gallotti[b] and C. Agostinelli[a]

[a] *Università di Trento* [b] *Fondazione Bruno Kessler (Trento, IT)*

Emerged as one of modern society's most pressing challenges, disinformation campaigns are orchestrated to distort facts and sow societal discord. The manipulative power of these campaigns have been significantly enhanced by *coordinated inauthentic behaviours* (CIBs) - disguised online groups coordinated in content and timing to spread disinformation. This research investigates the detection of CIBs focusing on their anomalous temporal patterns.

We model social media activities via a mixture of Hawkes processes (HPs), stochastic processes that capture event timings and self-exciting interactions. Building on event sequence clustering methodologies [1], we opted for a Bayesian hierarchical approach using a Dirichlet distribution as the prior distribution for the K mixture components, an exponential function controlling users' influences ϕ, and exponential priors for each component's parameters (μ_k, α_k, β_k). We proceed inferring parameters and latent clustering via a Metropolis Hastings within Gibbs algorithm.

Addressing the challenge of limited ground-truth data, we first evaluate our model via a simulation framework leveraging HPs. Carefully selecting parameters, we generate synthetic CIBs mimicking different strategies: from spammers to sockpuppets (i.e. users intensely interacting to disguise themself). Our findings highlight the effectiveness of our model, obtaining high homogeneity and completeness. To further evaluate our model, we are currently experimenting on a labelled high-dimensional dataset of CIBs. Our preliminary results confirm the efficacy of our model in identifying real CIBs, paving the way to more accurate and adaptable strategies against disinformation.

Keywords: Disinformation, Mixture models, Hawkes processes

References

[1] Xu, H., & Zha, H. (2017). A dirichlet mixture model of hawkes processes for event sequence clustering. *Advances in neural information processing systems*, **30**.

Analysing the efficiency of deseasonalisation methods in removing seasonal bias in outliers

R. Netshiomvani[a], S. C. Liebenberg[a] and G. L. Grobler[b]

[a]*Focus Area for Pure and Applied Analytics, North-West University* [b] *Unit for Data Science and Computing, North-West University*

Deseasonalisation is a common preprocessing step in climate time series analysis, particularly when the focus is anomaly detection, where seasonal patterns can mask or bias extreme events. Most deseasonalisation approaches focus on removing the seasonal mean, implicitly assuming that this is sufficient to eliminate seasonal patterns. However, many meteorological variables exhibit seasonality in both mean and variance, which may bias the detection of outliers. In this talk we will show whether deseasonalisation methods effectively remove seasonal bias in outliers from bivariate climate time series obtained from weather station and ERA5 data. Daily wind speed data from the South African Lowveld are analysed using paired observations from both sources. Standard deseasonalisation methods are used to remove seasonality in the mean, while less well-known approaches also account for seasonal variance. In addition, the effects of serial correlation and conditional heteroscedasticity are removed by applying an ARMA–GARCH model and analysing the resulting standardised residuals. To assess the presence of seasonal bias, the monthly distribution of detected outliers is evaluated using a chi-square goodness of fit test. The results suggest that reliable anomaly detection in bivariate climate data may require deseasonalisation methods that account for seasonality in both the mean and the variance.

Keywords: Deseasonalisation, Seasonal bias, Wind speed.

Polynomial-time near-optimal estimation over certain type-2 convex bodies

M. Neykov

Northwestern University, USA

We will discuss near (minimax) optimal computationally efficient estimators for constrained estimation over certain special convex constraints. The methodology is applicable to the problem of the constrained Gaussian sequence model, as well as constrained robust mean estimation and regression. Examples of sets over which this method can be applied include all well-balanced linearly transformed quadratically convex orthosymmetric sets.

Keywords: Constrained estimation, robust estimation.

Sparse Feature Group K-Means

A. W. Diallo[a], M. Ouattara[b], N. Niang[c] and M. Bousso[d]
[a],[d] *Université Iba Der Thiam, Thies, Senegal*
[b] *Université Polytechnique de San Pedro, San Pedro, Côte d'Ivoire*
[c] *CEDRIC – CNAM, Paris, France*

We address the problem of clustering high-dimensional multiblock data, where features are organized into homogeneous blocks representing different views of the data. Traditional subspace clustering methods like Entropy Weighting K-Means (EWKM) [1] and Feature Group K-Means (FGKM) [2] assign continuous weights to features or feature blocks, requiring tedious post-hoc analysis to identify cluster-relevant elements based on weight magnitudes. Sparse clustering methods such as Sparse K-Means [3] and Sparse Subspace K-Means [4] offer automatic feature selection but do not exploit the block structure inherent in multiblock data.

To address this limitation, we propose SFGKM (Sparse Feature Group K-Means), which extends Sparse Subspace K-Means (SSKM) to incorporate multiblock data structure. SFGKM employs a unified optimization criterion with dual-level lasso-type penalties that simultaneously performs observation clustering, cluster-specific individual feature selection, and cluster-specific feature block selection. The method automatically sets irrelevant feature and block weights to zero while assigning large weights to cluster-characterizing elements, eliminating manual weight analysis.

The optimization follows an alternating maximization scheme that iteratively updates cluster assignments, feature weights within blocks, and block weights using soft-thresholding operators until convergence. Experimental validation is conducted on synthetic datasets with varying complexity and noise patterns, as well as on real-world multiblock datasets covering diverse domains and block structures. Results show competitive performance on synthetic data, where SFGKM correctly identifies cluster-specific relevant blocks and automatically zeros out noise features. On real-world data, SFGKM achieves superior clustering performance compared to existing methods, demonstrating that explicitly modeling block structure is essential when views must be treated as coherent units rather than collections of independent features.

Keywords: Sparse Clustering, Soft Subspace Clustering, Feature Selection, Multi-block data.

References

[1] L. Jing, M. Ng, J. Huang (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, 1026–1041.

[2] X. Chen, Y. Ye, X. Xu, J. Z. Huang (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 434–446.

[3] D. M. Witten, R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 713–726.

[4] A. W. Diallo, N. Niang, M. Ouattara (2021). Sparse subspace k-means. In *ICDMW 2021*, 678–685.

Sparse Weighted Multi-view Clustering

N. Niang[a], M. L. Ndao[a] and M. Ouattara[b]

[a] *Conservatoire National des Arts et Métiers CNAM, Paris* [b] *Université San Pédro, Côte d'Ivoire*

We address multi-view clustering in which individuals are described by variables that are partitioned into several homogeneous and meaningful blocks. As the blocks are intended to be homogeneous, preserving this block homogeneity would help reveal the underlying structure of the individuals. So at a first level, the individuals are clustered according to each block separately, and the resulting partitions (contributory partitions) are aggregated in a consensual partition in a second step [4]. Therefore, this multi-view clustering issue is reformulated as a consensus of partitions problem. The choice of the first step clustering method is not addressed here. We only focus on the aggregation of the obtained partitions. These partitions are seen as categorical variables and then associated with indicator matrices and connectivity matrices whose entries are 1 if two individuals are in the same cluster and 0 if not. Using connectivity matrices avoids the label switching issue. It has been pointed out that simple consensus methods, such as CSPA (Cluster based Similarity Partitioning Algorithm) [6], can yield unstable results when the contributory partitions are significantly different and if some of them are highly correlated.This redundancy could bias the final partition towards these correlated partitions. To address these limitations, weighted consensus methods have then been proposed with methods such as Weighted Non Matrix Factorization (WNMF) [3]. We propose a sparse weighted consensus method based on Constrained Singular Value Decomposition [1] and the RV correlation coefficient [5] between the connectivity matrices to find an unique partition from contributory ones. The results on simulated data as well as real ones show the relevance of the proposed method particularly when dealing with redundant partitions. In addition, the RV-based STATIS [2] method allows visualisation of the multi-view data as well as the clustering results.

Keywords: Sparsity ; Multi-view Clustering ; RV Coefficient

References

[1] Vincent Guillemot, Derek Beaton, Arnaud Gloaguen, Tommy Löfstedt, Brian Levine, Nicolas Raymond, Arthur Tenenhaus, and Hervé Abdi. A constrained singular value decomposition method that integrates sparsity and orthogonality. *PloS one*, 14(3):e0211463, 2019.

[2] Christine Lavit, Yves Escoufier, Robert Sabatier, and Pierre Traissac. The act (statis method). *Computational Statistics & Data Analysis*, 18 (1):97–119, 1994.

[3] Tao Li and Chris Ding. Weighted consensus clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 798–809. SIAM, 2008.

[4] Ndèye Niang and Mory Ouattara. Weighted consensus clustering for multiblock data. In *Actes SFC 2019*, Paris, France, September 2019. URL https://cnam.hal.science/hal-02471611.

[5] Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3):257–265, 1976.

[6] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

interlace: Closing the Mixed-Effects Modelling Gap in Python

Helio Pais[a]

[a] *The StepStone Group*

Mixed-effects models are central to applied statistics, yet Python's ecosystem for fitting them remains limited compared to R. The most widely used Python option, statsmodels' MixedLM [3], cannot accommodate crossed random effects and offers no influence diagnostics beyond residuals. Practitioners who need these capabilities must either bridge to R via rpy2 (incurring computational overhead) or turn to Bayesian frameworks whose cost can be orders of magnitude higher. This leaves a large class of routine mixed-model analyses inaccessible from a Python workflow.

We present **interlace**, an open-source Python library for mixed-effects modelling. interlace implements profiled restricted maximum likelihood (REML) estimation with sparse Cholesky factorisation—the same algorithmic strategy underlying R's lme4 [1]—and supports both crossed and nested random intercepts and slopes through a formula interface. Beyond point estimation, the library provides estimated marginal means, Type I–III ANOVA tables, profile likelihood confidence intervals, parametric bootstrap, and group-level cross-validation—none of which are available together in any existing Python package.

interlace also provides regression diagnostics for hierarchical models: leverage, Cook's distance, MDFFITS, and COVTRACE at both the observation and group level, mirroring R's HLMdiag package [2]. These measures allow practitioners to assess the influence of individual observations and clusters on fixed-effect estimates and variance components—functionality that has been largely absent from the Python ecosystem.

Systematic benchmarks against lme4 confirm numerical parity: fixed-effect estimates agree within 10^{-4}. Computationally, interlace is competitive with or faster than lme4 on problems involving nested random effects (approximately $8\times$ faster at 10^6 observations) and operates within the same order of magnitude as fixest [4] on high-dimensional fixed-effects specifications. interlace is available on PyPI and GitHub under a BSD-3 licence.

Keywords: Mixed-effects models, REML estimation, regression diagnostics.

References

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

[2] A. Loy and H. Hofmann (2014). HLMdiag: A Suite of Diagnostics for Hierarchical Linear Models in R. *Journal of Statistical Software*, **56**(5), 1–28.

[3] S. Seabold and J. Perktold (2010). statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, 92–96.

[4] L. Bergé (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package fixest. *CREA Discussion Papers*, 13.

A robust and flexible nonparametric approach for longitudinal microbiome data analysis

Md. M. Hosain[a], H. Koh[b], and T. Park[a]

[a]*Department of Statistics, Seoul National University, Seoul, Korea,* [b]*Department of Applied Mathematics and Statistics, The State University of New York, Korea (SUNY Korea), Incheon, Korea*

Recent improvements and affordable sequencing technology have made it possible to conduct metagenomic studies on associations between members of microbial community with human health or disease in depth.

Longitudinal studies are essential to understand the dynamic relationship between microbiota and host phenotypes over time. However, most existing methods suffer from the inherent complexity, sparsity, over-dispersion, and high-dimensionality of longitudinal microbiome data, and therefore tend to address some but not all critical data characteristics. In this study, we propose the microbiome residual permutation and power transformation (MiRP) framework and two omnibus tests, including MiRP-O and MiRP-WO for association testing in longitudinal microbiome studies. The proposed methods employ generalized estimating equations (GEE), residual permutations, and power transformation of the microbial taxa to account for covariate effects and longitudinal correlation. MiRP-O uses the minimum p-value across various power-transformed abundances, whereas MiRP-W-O employs weighted averaging to enhance robustness.

We applied our proposed methods, MiRP-O and MiRP-W-O, to a simulation study and real data analyses. Through simulation studies, we showed that both MiRP-O and MiRP-W-O well control the Type I error rates and have higher power compared to other existing methods. Real data applications showed that methods were able to capture important microbial taxa related to phenotypes, indicating the effectiveness of the methods. MiRP-O and MiRP-W-O are both robust and powerful for capturing phenotype-related microbial markers in longitudinal microbiome studies.

Keywords: Microbiome, Longitudinal data, Relative abundance, Differential abundance analysis, Permutation test.

Selection Confidence Sets for Equally Weighted Portfolios

D. Ferrari[a], A. Fulci[b] and S. Paterlini[b]

[a] *Faculty of Economics and Management, Free University of Bozen-Bolzano,* [b] *Department of Economics and Management, University of Trento*

Given a universe of N assets, investors often form equally weighted portfolios (EWPs) by selecting subsets of assets. EWPs are simple, robust, and competitive out-of-sample, yet the uncertainty about which subset truly performs best is largely ignored. Traditional approaches typically rely on a single selected portfolio, thus failing to consider alternative investment strategies that may perform just as well when accounting for statistical uncertainty or model instability. To address this selection uncertainty, we introduce the Selection Confidence Set (SCS) for EWPs: the set of all portfolios that, under a given loss function and at a specified confidence level, contain the unknown set of optimal portfolios under repeated sampling. The SCS quantifies selection uncertainty by identifying a range of plausible portfolios, challenging the idea of a uniquely optimal choice. Like a confidence set, its size reflects uncertainty – growing with noisy or limited data, and shrinking as the sample size increases. Theoretically, we show that the SCS achieves asymptotic coverage of any fixed population-optimal selection and characterize how its size depends on underlying uncertainty, corroborating these findings through Monte Carlo experiments. Applications to the French 17-Industry Portfolio and Layer-1 Cryptocurrency data underscore the importance of accounting for selection uncertainty when comparing equally weighted strategies.

Keywords: Equally Weighted Portfolios, Selection Confidence Set, Selection Uncertainty, Subset Selection, Wald Test.

Pairwise Fisher transformation of a conditional correlation matrix

C. Francq[a], S. Laurent[b], M. Plazzogna[c] and J-M. Zakoian[a]

[a] *CREST-ENSAE and Lille University,* [b] *Aix-Marseille University,* [c] *University of Geneva*

This paper introduces a novel Multivariate GARCH (MGARCH) framework for modeling dynamic conditional correlation matrices using a generalized pairwise Fisher transformation. We propose a flexible multivariate volatility model where the conditional correlation between any two assets is governed by a stochastic recurrence equation mapped through a known bijection to the interval $(-1, 1)$.

We establish the theoretical conditions necessary for the existence of a unique, strictly stationary, and ergodic solution for both bivariate and general N-variate specifications. To tackle the curse of dimensionality inherent in MGARCH models, we develop a multi-step quasi-maximum likelihood estimation (QMLE) procedure that estimates GARCH-type volatilities equation-by-equation and correlations pair-by-pair. We formally derive the strong consistency and asymptotic normality of these estimators.

Furthermore, because the independent assembly of pairwise correlations does not inherently guarantee a positive definite global correlation matrix, we implement an ex-post metric projection step. Crucially, we prove that the entrywise perturbation introduced by this nearest-positive-definite projection is bounded by the asymptotic estimation error, thereby preserving the $\sqrt{T}$-consistency of the initial multi-step estimators.

Keywords: MGARCH, Conditional Correlation, Multi-step QMLE.

Hyperevent network modelling of partially observed gossip data

V. Poda[a,b], V. Vinciotti[a] and E. C. Wit[c]

[a]*Department of Mathematics, University of Trento, Trento, Italy,* [b]*Fondazione Bruno Kessler, Trento, Italy,* [c]*Università della Svizzera italiana, Lugano, Switzerland*

Gossiping is a widespread social phenomenon that shapes relationships and information flow in communities. From a network theoretic point of view, gossiping can be seen as a higher-order interaction, as it involves at least two persons talking about a non-present third. The mechanism of gossiping is complex: it is most likely dynamic, as its intensity changes over time, and possibly viral, if a gossiping event induces future gossiping, such as a repetition or retaliation. We define covariates of interest for these effects and propose a relational hyperevent model to study and quantify these complex dynamics. We consider survey data collected yearly from 44 secondary schools in Hungary. No information is available about the exact timing of the events nor about the aggregate number of events within the yearly time interval. What is measured is whether at least one gossiping event has occurred in a given time interval. We extend inference for relational hyperevent models to the case of right-censored interval-time data and show how flexible and efficient generalized additive models can be used for estimation of effects of interest. Our analysis on the school data illustrates how a model that accounts for linear, smooth and random effects can identify the social drivers of gossiping, while revealing complex temporal dynamics.

Keywords: dynamic network, right-censoring, relational hyperevent model

References

[1] V. Poda, V. Vinciotti, and E. C. Wit (2025). *Hyperevent network modelling of partially observed gossip data*. arXiv:2511.18543. `https://arxiv.org/abs/2511.18543`

Cellwise robust Gaussian mixture model for multi-group data with label noise

P. Puchhammer[a], I. Wilms[b] and P. Filzmoser[a]

[a] *Technische Universität Wien,* [b] *Maastricht University*

Do expert-defined or diagnostically-labeled data groups align with clusters inferred through statistical modeling? If not, where do discrepancies between predefined labels and model-based groupings occur and why? We introduce the multi-group Gaussian mixture model (MG-GMM), the first model developed to investigate these questions. It incorporates prior group information while allowing flexibility to reassign observations to alternative groups based on data-driven evidence.

To this end we model the data based on Gaussian Mixture Models. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ be given data sets from N groups consisting of $\boldsymbol{X}_g = ((\boldsymbol{x}_{g,1})', \ldots, (\boldsymbol{x}_{g,n_g})')' \in \mathbb{R}^{n_g \times p}$ independent observations, for $g = 1, \ldots, N$, of the same p variables. Assume that each observation $\boldsymbol{x}_{g,i}$ from group g originates from a Gaussian mixture

$$\boldsymbol{x}_{g,i} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \text{ with probability } \pi_{g,k} \geq 0$$

for $k = 1, \ldots, N$, and where $\varphi(\boldsymbol{x}_{g,i}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is defined as the multivariate normal density for $\boldsymbol{x}_{g,i}$. Based on the assumption that each individual group is coherent, assume $\pi_{g,g} \geq 0.5$. Thus each group g has a main distribution $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. However, data-driven reassignment of observations outlying in the original group is allowed by the flexibility of the mixture model.

Moreover, our model offers robustness against cellwise outliers that may obscure or distort the underlying group structure based on a penalized likelihood approach. The proposed methodology implemented via an EM-type algorithm provides good simulation results and its potential is illustrated on wine data.

Acknowledgements: Co-funded by the European Union (SEMACRET, Grant Agreement no. 101057741) and UKRI (UK Research and Innovation). Ines Wilms is supported by a grant from the Dutch Research Council (NWO),

research program Vidi under the grant number VI.Vidi.211.032.

Keywords: Gaussian mixture models, cellwise outliers, labeled data

Compositional microbiome-based signatures associate with general health status

M. Pujolassos[a], A. Kurilshikov[b], S. Zhernakova[b] and M.L. Calle[c]

[a] *University of Vic – Central University of Catalonia, Spain,* [b] *University Medical Center Groningen, Groningen, The Netherlands,* [c] *Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central, Spain*

Growing evidence links gut microbiome composition to human health. However, many disease-associated microbial signatures come mainly from clinical cohorts, leaving their relevance to general-population health unclear. This gap, together with the challenges of analysing compositional microbiome data, still limits their use in preventive medicine.

To address this challenge, we adopted a two-stage aproach combining (1) discovery of disease-associated microbial signatures from public datasets and (2) external validation in a general population cohort. Bacterial signatures were identified using a compositional data analysis method, coda4microbiome [2], which performs variable selection via penalized regression applied to all-pairs of log-ratios. Using coda4microbiome in large public microbiome datasets, we identified 16 disease-specific bacterial signatures, which were then quantified in a general-population cohort, LifeLines-DMP [3], to assess their association with overall health status and well-being.

Most signatures showed consistent associations linked to poorer overall health. Results support the idea that shifts in microbial compsition may reflect early physiological alterations preceding clinical diagnosis, highlighting the value of microbiome-based compositional biomarkers in preventive healthcare.

Keywords: compositional data analysis, microbiome, microbial signatures, general health prediction

References

[1] M.L. Calle, et al. (2023) coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinformatics*, **24**

[2] R. Gacesa, et al. (2022) Environmental factors shaping the gut microbiome in a Dutch population. *Nature*, **604**, 732–739.

Multivariate functional Mahalanobis distance with application to clustering

U. Radojicic[a], J. Oguamalam[a], T. Masak[b] and P. Filzmoser[a]

[a] *TU Wien, Vienna, Austria,* [b] *WU Wien, Vienna, Austria*

The increasing availability of multivariate functional data across diverse scientific domains highlights the importance of precise analyses of such data structures. Unlike univariate functional data, multivariate functional observations not only exhibit temporal dependence but also between-component correlation. A common simplifying assumption, used to make the estimation of the covariance more tractable, is the separability of the covariance operator, which assumes that the correlation across time and between the components can be uncoupled.

Recently, the α-Mahalanobis distance was introduced to the univariate functional setting. Building on this idea, this work introduces the regularized multivariate Mahalanobis distance (RMMD) as an extension of this metric to the multivariate functional case under a separable covariance structure. The incorporation of an appropriately chosen regularization operator assures that the RMMD of a multivariate stochastic process can be calculated as the sum of univariate α-Mahalanobis distances of its scaled and decorrelated components.

We demonstrate the utility of the RMMD in the context of distance-based clustering of multivariate functional data.

Keywords: multivariate functional data, separable covariance, distance-based clustering

References

[1] McCormack, A., & Hoff, P. (2025). *Information geometry and asymptotics for Kronecker covariances.* Bernoulli, **31**(4), 3165-3186.

[2] Oguamalam, J., Radojičić, U., & Filzmoser, P. (2024). *Minimum regularized covariance trace estimator and outlier detection for functional data.* Technometrics, **66**(4), 588–599.

Cellwise outliers: from identification to regression

J. Raymaekers[a] and P.J. Rousseeuw[b]

[a] *University of Antwerp, Belgium,* [b] *KU Leuven, Belgium*

Modern datasets often contain cellwise outliers, where some individual entries of the data matrix are contaminated. This distinction is important because unaffected cells in a contaminated row may still contain valuable information. We briefly discuss the cellHandler method for identifying cellwise outliers. We explain how it led to the development of cellMCD, a cellwise robust extension of the Minimum Covariance Determinant estimator, for estimating a location and scatter matrix under cellwise contamination. The main focus of the presentation is a new cellwise robust regression methodology called cellLTS. The method achieves the first breakdown result for cellwise robust regression and is specifically designed to provide reliable out-of-sample predictions. A real-data example of modelling cancer rates in counties in the US illustrates the capabilities of cellLTS of providing fresh insights.

Keywords: Cellwise outliers, MCD estimator.

Scalable Inference for Individual-Based Epidemic Models

Lorenzo Rimella

Università degli studi di Bergamo

Individual-based models allow epidemiologists to capture risk at the individual level. However, the computational cost of exact likelihood evaluation for partially observed individual-based models grows exponentially with population size, necessitating approximate inference. In this contribution, we explore scalable approximations to the likelihood for individual-based models that avoid the exponential-in-population computational cost. In particular, we analyse two approaches: Simulation-Based Composite Likelihood [1], which uses a composite likelihood approximation, and Categorical Approximate Likelihood [2], which constructs an approximation by substituting expectations in the same vein as assumed density filters. This work analyses the advantages and disadvantages of both methods and further compares them with block particle filters, while also outlining directions for future research.

Keywords: Epidemiology, Approximate Inference, Hidden Markov Models

References

[1] Rimella, Lorenzo and Jewell, Chris and Fearnhead, Paul (2025). Simulation based composite likelihood. *Statistics and Computing*, **35**(3), Springer.

[2] Rimella, Lorenzo and Whiteley, Nick and Jewell, Chris and Fearnhead, Paul and Whitehouse, Michael (2026). Scalable calibration of individual-based epidemic models through categorical approximations. *Journal of the American Statistical Association*, to appear.

Optimal Self-Distillation in Ridge Regression: Sharp Asymptotics and One-Shot Tuning

D. Hien[a], P. Patil[a] and A. Rinaldo[a]

[a] *The University of Texs at Austin*

Self-distillation (SD) is the process of retraining a student model on a mixture of ground-truth labels and a teacher models predictions, using the same architecture and training data. While SD has been empirically shown to improve generalization in regression tasks, a rigorous theoretical under- standing of its mechanics and properties remains elusive. We study SD for ridge regression in the general unconstrained setting where the mixing weight is allowed to lie outside the unit interval. Without any distributional assumptions, we prove that the squared prediction risk including the out-of-distribution risk of the optimally mixed student model strictly improves upon the teacher model for every value of regularization at which the teacher's risk is non-stationary. We express the optimal mixing weight in terms of the teacher's risk derivative, thereby characterizing the somewhat surprising scenarios in which a negative weight is optimal. To quantify the magnitude of these improvements, we derive exact risk asymptotics in the proportional regime under general anisotropic covariances and deterministic signals. Building on this theory, we propose a novel one- shot tuning procedure to consistently estimate the optimal mixing weight without retraining, sample splitting, or grid search. Experiments on real-world regression tasks and pre-trained neural net- work features validate our theoretical predictions and demonstrate the effec- tiveness of the proposed tuning methodology.

Keywords: Self-Distillation, Ridge Regression, Tuning.

Beyond the smoke: What Statistics reveal about Brazilian wildfires

P.C. Rodrigues[a,b], J. Pimentel[c], R. Bulhões[a] and A. Pinheiro[a]

[a] *Department of Statistics, Federal University of Bahia, Salvador, Brazil* [b] *Department of Business Management, University of Pretoria, Pretoria, Pretoria, South Africa,* [c] *Department of Statistics, Federal university of Pernambuco, Recife, Brazil*

Wildfires are among the most common natural disasters in many world regions and actively impact life quality. These events have become frequent due to climate change, other local policies, and human behaviour. In this talk, I consider the historical data with the geographical locations of all the "fire spots" detected by the reference satellites covering the Brazilian territory between January 2011 and December 2022, comprising more than 2.2 million fire spots. First, I will show the results of a spatio-temporal generalized linear model for areal unit data, whose inferences about its parameters are made in a Bayesian approach and use meteorological variables (precipitation, air temperature, humidity, and wind speed) and a human variable (land-use transition and occupation) as covariates. Then, I will present the results for the hierarchical time series forecasting where the six Brazilian biomes and the 5570 municipalities form the hierarchy.

Keywords: Brazilian wildfires, Spatio-temporal modeling, Hierarchical time series.

References

[1] A.C. Pinheiro, and P.C. Rodrigues (2024). Hierarchical time series forecasting of fire spots in Brazil: A comprehensive approach. *Stats*, **7**, 647–670.

[2] J. Pimentel, R. Bulhões, and Rodrigues, P.C. (2025). Beyond the Smoke: What Statistics Reveal About Brazilian Wildfires. *ISI Magazine*, **1**, 3–6.

[3] J. Pimentel, R. Bulhões, and P.C. Rodrigues (2024). Bayesian spatio-temporal modeling of the Brazilian fire spots between 2011 and 2022. *Scientific Reports*, **14**, 21616.

Structural equation modeling with instrumental variables

Y. Rosseel[a]

[a] *Ghent University*

Instrumental variables (IV) play a central role in econometrics and related disciplines, where they are widely used to address issues of endogeneity and causal identification. In contrast, their use in psychometrics has remained limited. Early contributions considered IV estimation in the context of the factor model, but a key development was the seminal work of [1], which demonstrated how instrumental variables can be employed to estimate directed parameters–such as factor loadings and path coefficients–within structural equation models (SEMs). Nevertheless, nearly three decades later, IV methods are still rarely used in SEM and psychometrics more generally.

In this presentation, I review recent methodological advances that may facilitate a wider adoption of IV estimation in SEM. First, after using IV estimation for the directed parameters in the model, noniterative methods can now be used to estimate the remaining undirected parameters (variances and covariances) in a second stage. Moreover, analytic standard errors can be obtained for all free parameters in the model. Second, the availability of complete parameter estimates allows the derivation of model-implied moments, which in turn enables formal goodness-of-fit testing. Third, the theoretical relationship between traditional IV estimation and SEM has become clearer; for example, Browne's residual test can be shown to correspond to the Sargan test of overidentifying restrictions. Finally, IV estimation has recently been implemented in the `lavaan` package in R, further lowering the barrier to practical application.

Keywords: Structural Equation Modeling, Instrumental Variables.

References

[1] Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, **61**(1), 109–121.

The bixplot: A variation on the boxplot suited for bimodal data

C. M. Montalcini[a] and P. J. Rousseeuw[b]

[a] *Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland,* [b] *University of Leuven, Belgium*

Boxplots and related visualization methods are widely used exploratory tools for taking a first look at collections of univariate variables. In this note an extension is provided that is specifically designed to detect and display bimodality and multimodality when the data warrant it. For this purpose a univariate clustering method is constructed that ensures contiguous clusters (meaning that no cluster has members inside another cluster), and such that each cluster contains at least a given number of unique members. The resulting bixplot display facilitates the identification and interpretation of potentially meaningful subgroups underlying the data. The bixplot also displays the individual data values, which can draw attention to isolated points. Implementations of the bixplot are available in both Python and R, and their many options are illustrated on several real datasets. For instance, an external variable can be visualized by color gradations inside the display. For a full text with many figures see [1].

Keywords: Clustering, Graphical display, Violin plot.

References

[1] C. M. Montalcini and P. J. Rousseeuw (2025). The bixplot: A variation on the boxplot suited for bimodal data. *ArXiv report* 2510.09276, https://arxiv.org/abs/2510.09276.

Principal component analysis for interval-valued data with multiple number of observations from each subject

A. Roy

The University of Texas at San Antonio

We propose a novel method to obtain principal components (PCs) by using patterned variance-covariance matrix that partitions the actual variance-covariance matrix into between subject variation, within interval variation, and within multiple observations variation of the data if the data have multiple observations from each subject ([2]). We apply our method to a Face dataset (Table 1, [1]) that was obtained from a study of face recognition patterns for surveillance purposes: sequence of images (video frame from video source) were obtained with six features and with three sequences from each face. The first PC represents the 'landmark triangle' specified by three features (do not move) to quantify a face and the second PC represents the 'movable portions' specified by the other three features of a face.

We study a 'circle of correlation plot' that adroitly gives a quick and nice visual interpretation on how features are correlated with the PCs. 'Circle of correlation plot' clearly show the features related to 'landmark triangle' contribute mostly to PC1, while features related to 'movable portions' contribute mostly to PC2.

We presented a comparison study showing the correlation of the original variables and the PC1 generated by some previous methods and our proposed method. We see the component correlations from our method are mostly stronger or as good as in absolute value in comparison to previous methods.

Keywords: Interval-valued data, Circle of correlation plot.

References

[1] A. Douzal-Chouakria, L. Billard, and E. Diday (2011). Principal component analysis for interval-valued observations. *Stat Anal Data Min* **4**(2), 229–246.

[2] A. Roy (2025). Two-stage principal component analysis on interval-valued data using patterned covariance structures. *Adv Data Anal Classif*, https://doi.org/10.1007/s11634-025-00650-9

Hellinger-Type Losses for Generative Adversarial Networks

G. Saraceno[a], A. N. Vidyashankar[b] and C. Agostinelli[c]

[a]*Department of Statistical Sciences, University of Padua, Italy,* [b]*Department of Statistics, George Mason University, USA,* [c]*Department of Mathematics, University of Trento, Italy*

Generative adversarial networks (GANs) are commonly studied through the convergence of the induced distributions, while less attention has been given to the joint statistical behavior of the generator and discriminator estimators, and to their robustness under contamination. In this study, we investigate GAN training from a statistical perspective by introducing Hellinger-type adversarial loss functions, motivated by the boundedness and symmetry of the Hellinger distance.

Within a parametric framework, we formulate adversarial training as a joint minimax estimation problem for the parameters of the generator and discriminator. We consider a complete Hellinger formulation and a tractable approximated version, and we study their statistical properties. In particular, we prove the existence and uniqueness of the minimax solution, and establish consistency and joint asymptotic normality. We also derive profiled asymptotic results for the generator parameter. To investigate robustness, we compute the influence functions, thereby linking adversarial learning to robustness theory and M-estimation. Finally, we also briefly discuss how the same construction can be extended to other f-divergence-based adversarial loss functions. The theoretical analysis is complemented by a simulation study in a Gaussian parametric setting, where we introduce controlled contamination to examine how different adversarial losses affect estimation accuracy and training dynamics. Additionally, we provide an illustration on a higher-dimensional image generation problem using the Fashion-MNIST dataset. Theoretical details and simulation results are available in [1].

Keywords: GANs, Robust statistics, Generative models.

References

[1] G. Saraceno, A. N. Vidyashankar, and C. Agostinelli (2025). *Hellinger loss function for Generative Adversarial Networks*. arXiv:2512.12267.

A State-Restricted Hidden Markov Model for Authorship Attribution of the Deutero-Pauline and Pastoral Epistles

Josiah Leinbach[a], Xuwen Zhu[b] and Shuchismita Sarkar[a]

[a]*Bowling Green State University,* [b]*The University of Alabama*

The New Testament contains thirteen epistles attributed to the Apostle Paul, all of which were traditionally accepted as authentically Pauline by early Christian theologians. Since the 19th century, however, many scholars have questioned Paul's authorship of certain epistles due to differences in vocabulary and writing style compared to the undisputed Pauline epistles. In particular, two clusters of epistles, known as the Deutero-Pauline Epistles (Ephesians, Colossians, and 2 Thessalonians) and the Pastoral Epistles (1 Timothy, 2 Timothy, and Titus) have been subject to the most doubt. This study presents a novel state-restricted hidden Markov model that constructs a constrained state space for the undisputed Pauline epistles and an unrestricted state space for other epistles. The model jointly analyzes all thirteen epistles and some additional biblical texts, employing a novel first-order Markov emission for transitions between parts of speech to classify sentences based on *Pauline* and *non-Pauline* style detection. Then, informed by New Testament scholarship, the result of the model is interpreted and the possibility of Pauline authorship for the Deutero-Pauline and Pastoral Epistles has been examined.

Keywords: hidden Markov model, stylometry, authorship attribution. (Use at most 3 keywords)

Seasonal Variability in Thermodynamic Conditions Preceding Heavy Rainfall in the Free State, South Africa

I. M. Schoeman and A. van den Heever

North-West University, Potchefstroom Campus, South Africa

Forecasting heavy rainfall in early austral summer (October to December) requires different techniques than in late austral summer (January to March). Hence, threshold values associated with heavy rainfall will also change over this period.

The aim of this presentation is to identify the thermodynamic variables that occur in conjunction with heavy rainfall events in the Free State province in South Africa and to determine the threshold values of these variables before heavy rainfall. The process involves establishing a baseline mean state of the applicable thermodynamic variables for each month during the study period and then compare this against the thermodynamic variable means observed on the specific days that met the criteria for heavy rainfall.

Improving understanding of thermodynamics in this data-scarce region—often called South Africa's breadbasket—could strengthen early warning systems. Earlier alerts for heavy rainfall would help mitigate potential impacts. This study is conducted using an event set comprising over 30 years of weather station and ERA5 data.

Keywords: Seasonality, Thermodynamics, Heavy rainfall.

References

[1] L. Dyson (2009). Heavy daily-rainfall characteristics over the Gauteng Province, *Water SA*, **35**(5), doi:10.4314/wsa.v35i5.49188.

[2] L.L. Dyson, L.L., J.H. Heerden, P. van and Sumner (2014). A baseline climatology of sounding-derived parameters associated with heavy rainfall over Gauteng, South Africa, *International Journal of Climatology*, **35**(1), doi:10.1002/joc.3967.

[3] O. J. Matthew, O. E. Abiye, and M. A. Ayoola (2021). Assessment of static stability indices and related thermodynamic parameters for predictions of atmospheric convective potential and precipitation over Nigeria, *Meteorology and Atmospheric Physics*, **133**(3), doi: 10.1007/s00703-020-00772-z.

[4] A. Bardóssy and G. Pegram (2016). Combination of radar and daily precipitation data to estimate meaningful sub-daily point precipitation extremes, *Journal of Hydrology*, **544**, doi: 10.1016/j.jhydrol.2016.11.039.

[5] F. M. Mashao et al. (2022). Extreme Rainfall and Flood Risk Prediction over the East Coast of South Africa, *Water*, **15**(1), doi: 10.3390/w15010050.

[6] M. G. Mengistu, C. Olivier, J. Botai, A. J. Adeola, and S. Daniel (2021). Spatial and temporal analysis of the mid-summer dry spells for the summer rainfall region of South Africa, *Water SA*, **47**(1), 76–87. doi: 10.17159/wsa/2021.v47.i1.9447.

Age-Specific Mortality in Italian Provinces via Functional Generalized Linear Mixed Models

M. Scianna[a] and C. Agostinelli[a]

[a] *Department of Mathematics, University of Trento*

Generalized Linear Mixed Models (GLMMs) are a standard tool for longitudinal and grouped data, yet their classical formulation assumes scalar predictors, discarding the granular information carried by covariates that are more naturally represented as functions. This work addresses a primary theoretical gap: the incorporation of distribution functions as functional covariates within a FGLMM framework. Treating predictors as distributions – rather than general smooth functions – allows one to assess how shifts in the mass of a population profile propagate into a scalar response, a setting not covered by existing scalar-on-function regression theory.

The proposed approach combines B-spline basis expansions with functional principal component analysis (FPCA) to reduce the infinite-dimensional covariate to a finite vector of uncorrelated scores [1]. These scores enter a hierarchical mixed-effects model with specific random intercepts and functional coefficients, estimated via both a Bayesian Hamiltonian Monte Carlo scheme [4] and a frequentist maximum likelihood baseline. The close agreement between the two provides a robustness check on the inferred functional effects. The framework is validated on ISTAT data covering 107 Italian provinces over 2011–2023, where age distributions serve as functional covariates and crude mortality rates as the response. Temporal regimes are identified non-parametrically through pairwise stochastic dominance tests [2], avoiding parametric assumptions on the response distribution. Reconstructed coefficients confirm a stable pre-pandemic age–mortality gradient, a marked amplification of middle-age contributions in 2020 consistent with the demographic profile of COVID-19 [3], and a progressive return to baseline through 2023. The application demonstrates that distributional covariates can be embedded in mixed models in a computationally feasible and interpretable way.

Keywords: Functional data analysis, Mortality modelling, Mixed models.

References

[1] J. O. Ramsay and B. W. Silverman (2005). *Functional Data Analysis*. Springer, New York.

[2] G. F. Barrett and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica*, **71**(1), 71–104.

[3] V. Kontis et al. (2020). Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. *Nature Medicine*, **26**(12), 1919–1928.

[4] O. Abril-Pla et al. (2023). PyMC: A modern and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, **9**(e1516).

s

Sensitivity analysis for causal effects measured on the Odds Ratio scale

E. Stanghellini[a] and S. Geneletti [b]

[a] *Dipartimento di Economia, Università di Perugia (IT)* [b] *Department of Statistics, London School of Economics and Political Science (U.K.)*

Causal conclusions drawn from observational studies necessarily rely on the unverifiable assumption that there are no unobserved confounders. With reference to the context of a binary treatment and a binary outcome, we propose two different - though related - strategies to assess sensitivity to a binary unobserved confounder when the causal effect is expressed as a (log) odds ratio, a situation commonly arising in standard logistic modelling, particularly in case-control studies.

Existing methods tipically rely on the rare outcome approximation. Using the exact formula linking marginal and conditional odds ratios in Stanghellini and Doretti (2019), we propose two graphical tools that visualize the extent to which the unobserved confounder could attenuate, nullify, or even reverse the estimated causal effect. Connections with Cornfield's conditions on relative risks are presented, thereby enlarging the circumstances where the proposed procedures can be applied.

The talk is based on a paper to appear (Stanghellini and Geneletti, 2026) and complements work for a continuous unobserved confounder (Gasparin et al., 2025).

Keywords: Cornfield's conditions, relative risk, visualization.

References

[1] Gasparin M., Scarpa B. and Stanghellini E. (2025). Omitting continuous covariates in binary regression models: Implications for sensitivity and mediation analysis. *Statistica Neerlandica*, **79**(1): e12369.

[2] Stanghellini E. and Geneletti S. (2026). Sensitivity analysis on the Odds Ratio scale: extending Cornfield's conditions. *Observational Studies*, to appear.

[3] Stanghellini E. and Doretti M. (2019). On marginal and conditional parameters in logistic regression. *Biometrika*, **106** (3):732–739.

Visualising education policy reforms and inequality indicators in Europe since 2000

G. Filandrianos[a], M. Symeonaki[b] and G. Stamou[a]

[a]*National Technical University of Athens,* [b]*Panteion University of Social and Political Sciences*

The increasing availability of large-scale comparative data on education systems creates new opportunities for analysing the relationship between policy reforms and educational assessment outcomes. However, these data are often dispersed across separate sources and presented independently, making it difficult for researchers and policymakers to explore potential connections between policy interventions and indicator trends. This paper presents the STRIDE interactive map, a visual analytics platform that integrates longitudinal education inequality indicators with a comprehensive database of education policy reforms across European countries. The map extends earlier indicator visualisation work on digital skills and attitudes towards ICT [3]. In the current study, education inequality indicators are estimated from raw microdata drawn from four large-scale international assessment surveys, i.e., PISA, ICILS, TIMSS, and PIRLS, and harmonised to produce comparable measures across countries and over time. Indicators are further disaggregated by key dimensions of inequality, including gender, socio-economic status, migration background, and school location [2]. In parallel, the platform incorporates a database of more than 430 education policy reforms implemented in European countries since the early 2000s [4, 1]. The platform is organised into two complementary visual environments: an indicators environment, presenting longitudinal trends in education indicators, and a reforms environment, providing an overview of policy activity across countries, years, education levels, evaluation status, coverage and reform types. The linkage between the two environments is implemented in the indicators interface, where policy reforms are visualised as temporally aligned events along indicator trends. Indicator pages display time-series data for a selected indicator for a specific country while simultaneously visualising reforms as typed events aligned with the temporal axis. Reforms are categorised by policy type and can be interactively explored through a details-on-demand interface providing contextual information about each reform. By combining

interactive mapping, temporal indicator visualisation and policy event annotation, the platform enables exploratory analysis of links between policy interventions and educational inequalities.

Keywords: Visual analytics, Social Sciences, Application Motivated Visualisation, Cartography and Maps, Visual Representation Design

Funding This work was conducted in course of the STRIDE project (Strategies for Achieving Equity and Inclusion in Education, Training and Learning in Democratic Europe), which is co-funded by the European Union (GA 101132339) and UK Research and Innovation (GA 10108849).*

References

[1] M. Ślusarczyk, S. Nørgaard Iversen, M. Świkatkiewicz-Mośny, R. Sommer Hansen, and J. Michcik. *25 Years of Education Policy Changes for Equity and Inclusion in Europe – Policy Analysis Report*. Jagiellonian University and VIA University College, 2026. https://doi.org/10.5281/zenodo.17290758.

[2] I. Steinmann, Z. Tomka, A. Yfanti, and D. Bremer. *Key Indicators of Mapping Inequalities in Educational Achievements in Europe*. Lifelong Learning Platform, Brussels, Belgium, 2025. STRIDE Working Paper No. 1. Contributors: L. Huang, M. Symeonaki, I. Tóth. https://doi.org/10.5281/zenodo.17290711.

[3] M. Symeonaki, G. Filandrianos, and G. Stamou (2022). Visualising key information and communication technologies (ICT) indicators for children and young individuals in Europe. *Humanities and Social Sciences Communications*, 9, 351.

[4] M. Symeonaki, G. Filandrianos, and A. Gabós (2025). Data and methods of the interactive STRIDE map. *STRIDE Working Paper No. 2*. DOI: 10.5281/zenodo.17290753.

*The platform is available at https://apps.islab.ntua.gr/stride and will be fully functional from June 2026.

Recursive Computation of Multivariate Hermite–Gaussian Integrals with Applications

E. Taufer[a] and Gy. H. Terdik[b] and M. Bee[a]

[a] *University of Trento,* [b] *University of Debrecen*

An exact recursive algorithm for evaluating integrals of d-variate Hermite polynomials weighted by the Gaussian density over orthants is introduced. The proposed approach provides closed-form expressions that can be computed efficiently for any fixed polynomial order k and threshold vector. This result addresses a fundamental computational issue and enables the systematic use of higher-order expansions in practical multivariate problems.

The algorithm has a computational complexity of order $O(k\,d^k)$ and memory requirements of order $O(d^k)$, which may become prohibitive as either the dimension d or the Hermite order k increases. To mitigate this issue, a streaming formulation is developed that avoids storing large intermediate arrays arising from Kronecker products. This approach computes contributions sequentially, significantly reducing memory usage and improving numerical stability.

Applications are developed through integrated versions of Gram–Charlier and Edgeworth expansions for approximating multivariate cumulative distribution functions, as well as multivariate tail conditional expectations and related risk measures, including multivariate Value at Risk. The expansions are presented in compact and intuitive forms of arbitrary order using multivariate Bell polynomials.

Simulation studies and real data applications illustrate the performance and practical relevance of the proposed method. All results can be readily implemented using the `R` package `MultiStatM`. [1].

Keywords: Multivariate cumulants; multivariate density approximation; multivariate tail conditional expectation.

References

[1] G. H. Terdik and E. Taufer (2025). MultiStatM: Multivariate statistical methods in R. *The R Journal*, **16**(4):123—140.

Shrinkage and visualization of multivariate Gram-Charlier approximations

Gy. H. Terdik[a], and E. Taufer[b]

[a] *Universityof Debrecen,* [b] *University of Trento*

For univariate densities, the fourth-order Gram Charlier (GC) approximation is expressed through skewness- and kurtosis-driven Hermite corrections to a Gaussian baseline. Since truncation may produce negative values even when the true cumulants are used, ensuring positivity requires that the GC correction factor remain non-negative on a prescribed central interval. To address violations, we consider two shrinkage strategies: a one-parameter scaling of the full correction term, and a two-parameter scaling that separately shrinks the skewness and kurtosis components. The latter is formulated as a convex quadratic projection problem over a feasible set defined by linear inequalities on a dense grid, yielding a unique solution.

The idea is extended to the multivariate setting. Three nested correction schemes are proposed: global one-parameter scaling, K-parameter scaling by Hermite order, and a finest-grain distinct-parameter scaling acting on the distinct tensor coefficients of multivariate Hermite terms. By rewriting the multivariate correction in terms of distinct tensor entries and associated weights, the positivity constraints become linear, and the resulting shrinkage problem can be solved efficiently by convex optimization tools such as CVXR ([1]).

In addition, the work outlines visualization strategies for comparing multivariate densities and their approximations, including spiral-based representations in two dimensions and cone-based restrictions in three dimensions. These provide interpretable geometric views of fitted densities and help assess the effects of shrinkage in practice.

Keywords: Multivariate density estimation; convex optimization; geometric visualization.

References

[1] A. Fu, B. Narasimhan, S. Boyd (2020). CVXR: An R Package for Disciplined Convex Optimization. *J. of Statistical Software*, **94**(14), 1—34.

Beyond pairwise: investigating higher-order statistical behaviour in network psychometrics.

N. Van Santen[a], Y. Rosseel[a] and D. Marinazzo[a]

[a] *Department of Data-analysis, Ghent University*

Networks have become a popular data-analytic tool for investigating item interactions in psychometric data [1]. This approach exemplifies a shift towards understanding the parts vs. whole contribution of items towards the construct under investigation, together with identifying central items or interactions for possible interventions. We believe that it is beneficial to extend this analysis beyond only pairwise statistics, since different items can become important at different orders [2]. By not restricting ourselves to pairwise associations, we can illuminate a richer landscape of behavioural pictures that can together inform possible underlying mechanisms. The focus shifts from an intractable network inference problem to shrinking the state space of possible mechanisms. We use higher-order information theoretical measures to reanalyze a dataset used to investigate the construct of empathy [3] and one used in personality research [4]. We compare our results with those drawn for correlation-based network psychometrics in the original publications.

Keywords: Network science, information theory, higher-order

References

[1] Borsboom, D., Deserno, M.K., Rhemtulla, M. et al. Network analysis of multivariate data in psychological science. Nat Rev Methods Primers 1, 58 (2021).

[2] Marinazzo, D., Van Roozendaal, J., Rosas, F.E. et al. An information-theoretic approach to build hypergraphs in psychometrics. Behav Res 56, 8057–8079 (2024).

[3] Giovanni Briganti et al. Network analysis of empathy items from the interpersonal reactivity index in 1973 young adults, Psychiatry Research, Volume 265, 2018, Pages 87-92.

[4] Giulio Costantini et al., State of the aRt personality research: A tutorial

on network analysis of personality data in R, Journal of Research in Personality, Volume 54, 2015, Pages 13-29.

The Eidos in the Echo: Finding a Blessing in LLM Hallucination

Sijian Wang[a]

[a] *Department of Statistics, Rutgers University, USA*

In the prevailing discourse on Large Language Models, "hallucination"—the generation of plausible but factually incorrect assertions—is universally diagnosed as a critical pathology to be excised through reinforcement learning and retrieval-augmented architectures. This talk proposes a radical reframing: that these errors are not merely noise, but "generative echoes" offering a unique window into the latent topology of the model's conceptual framework. By treating hallucinations as a feature rather than a bug, we apply two distinct philosophical lenses to extract high-value semantic data.

First, we operationalize Edmund Husserl's phenomenological method of eidetic variation. Just as Husserl imagined varying an object's attributes to determine which are accidental and which are essential, we analyze the specific substitutions an LLM makes when it confabulates. By observing which semantic vectors the model treats as interchangeable, we can distill the invariant essence—the "eidos"—of complex concepts, revealing how the model structures reality beyond mere keyword association.

Second, we adopt a Nietzschean perspectivism to interpret the multiplicity of these errors. Rather than seeking a single "ground truth," we map the diversity of hallucinations to trace the genealogical landscape of meaning. This approach exposes the hidden biases, historical associations, and cultural narratives that steer the model's "will to power," transforming hallucination from a reliability problem into a diagnostic tool for dataset sociology. Ultimately, this dual approach charts a path toward "Hallucination-Informed Interpretability," enabling us to architect AI that possesses not just factual rigidness, but profound conceptual nuance.

Keywords: LLM hallucination, Phenomenology, Interpretability.

Modelling and inference of relational events

E. C. Wit[a]

[a] *Faculty of Informatics, Università della Svizzera italiana, Lugano, CH*

When asked to imagine a social network, one typically envisions a static graph $G = (V, E)$, where edges $(i, j) \in E$ indicate relationships between nodes. Early statistical network models, developed in the 1970s and formalized in the 1980s through exponential random graph models (ERGMs), are inherently static. With the increasing availability of temporal network data in the 1990s, extensions such as temporal ERGMs and stochastic actor-oriented models (SAOMs) were proposed.

Building on the latter's shift toward event-based representations, relational event models (REMs), introduced in 2008, conceptualize network dynamics as a multivariate point process. Specifically, interactions are modeled as a counting process of events (i, j, t) occurring in continuous time, with conditional intensity

$$\lambda_{ij}(t \mid \mathcal{H}_t) = Y(t)\lambda_0(t) \exp\{\beta^\top x_{ij}(t, \mathcal{H}_t)\},$$

where $\mathcal{H}_t$ denotes the history of past events and $x_{ij}(\cdot)$ encodes time-varying covariates and endogenous network effects. This formulation connects directly to non-homogeneous Poisson processes and allows for fine-grained modeling of temporal dependence. REMs have proven highly flexible in applications ranging from ecological invasion processes and bike-sharing systems to communication networks and interbank lending. In this talk, I will describe recent methodological developments that improve estimation, scalability, and interpretability, making REMs increasingly practical for applied researchers.

Keywords: Networks, Counting process, Inference

References

[1] Boschi, M. and Wit, E. C. (2026). Goodness of fit in relational event models. *Statistics and Computing*, 36(1), 4.

[2] Filippi-Mazzola, E. and Wit, E. C. (2024). Modeling non-linear effects with neural networks in REMs. *Social Networks*, 79, 25–33.

[3] Bianchi, F., Filippi-Mazzola, E., Lomi, A. and Wit, E. C. (2024). Relational event modeling. *Ann. Rev. of Stat. and Its Application*, 11.

A Compositional Data Analysis Framework for Diagnosing LLM Reasoning over Time Series Anomalies

Elif Beyza Akyıldız[a], Mehmet Ali Erkan[a], and Ceylan Yozgatlıgil[a]

[a] *Middle East Technical University, Department of Statistics*

Large language models are increasingly applied to structured time series reasoning tasks, yet how they allocate attention across sensor channels and whether that allocation reflects prediction quality remains poorly understood. Applying standard multivariate analysis directly to this type of data breaks the assumptions of the simplex and can lead to misleading correlations. We propose a method that first transforms attention distributions using a centered log-ratio transformation and then applies principal component analysis (PCA) to map LLM attention distributions into a space that is easier to understand. Within this framework, biplot visualizations are used to jointly represent attention compositions and sensor contributions, enabling simultaneous interpretation of model behavior and variable influence. This provides an intuitive geometric view of how attention allocation aligns with anomaly detection performance. We evaluated a range of instruction tuned LLMs on the RATS-40K benchmark [1] across multiple sensor domains and anomaly types that treat each model's attention output as a compositional vector. Our method provides a simple and easy way to evaluate how reliable LLMs are in time series reasoning. It requires only attention outputs, generalizes across sensor domains and model architectures, and opens a new geometric lens for explainable AI in multivariate settings.

Keywords: Time Series Reasoning, Anomaly Detection, Biplot Visualization.

References

[1] Y. Yang, Z. Liu, L. Song, K. Ying, Z. Wang, T. Bamford, S. Vyetrenko, J. Bian, and Q. Wen (2025). TIME-RA: Towards Time Series Reasoning for Anomaly Diagnosis with LLM Feedback. *arXiv preprint*, arXiv:2507.15066.

Index