ISNPS
2024

6th International Symposium on Nonparametric Statistics

ISNPS 2024

June 25-29, 2024 - Braga, Portugal

CONTENTS

Welcome Note - Foreword from the Chairs of the Scientific Committee

ISNPS 2024 | 6th International Symposium on Nonparametric Statistics

Braga, June 25-29 2024

Following the successful ISNPS Conferences in Chalkidiki (Greece), Cádiz (Spain), Avignon (France), Salerno (Italy) and Paphos (Cyprus), we are pleased to welcome you to the *6th International Symposium on Nonparametric Statistics* in Braga (Portugal). The meeting is organized by the Centro de Matemática of the Universidade do Minho, and co-sponsored by the Institute of Mathematical Statistics and the Bernoulli Society for Mathematical Statistics and Probability.

The conference venue is Altice Forum Braga, a modern convention center situated near the center of the city of Braga, in North Portugal. Braga perfectly combines its two-thousand-year-old history with invigorating youth and vitality. It boasts one of the oldest Sacro-Montes in Europe, the renowned Bom Jesus, making it the flagship of the Minho region. Additionally, Braga is home to Portugal's oldest cathedral and the Monastery of Tibães, which holds great significance as the motherhouse of the Benedictines. Braga proudly embraces its Roman legacy and is often referred to as the *Portuguese Rome* due to its origins as the Roman city of Bracara Augusta. The city's distinctive churches, splendid 18th-century houses, gardens, parks, and leisure spaces pay homage to its historical roots. The monuments, museums, and churches that adorn Braga serve as a testament to its long and illustrious history, providing visitors with a splendid glimpse into the city's glorious past.

The *6th International Symposium on Nonparametric Statistics* puts together recent advances and trends in many areas of nonparametric statistics, including statistical learning, Bayesian nonparametrics, functional data analysis, high-dimensional data, goodness-of-fit, Survival Analysis, nonparametric econometrics, multiple testing or extremes. With 7 plenary talks, 239 invited talks organized in 60 invited paper sessions, 78 contributed talks, 20 poster presentations, and almost 400 participants from all over the world, the symposium is a perfect place to facilitate the exchange of research ideas, promote collaborations and contribute to the further development of the field of nonparametric statistics. We want to thank all the colleagues in the ISNPS 2024 Scientific Committee and the invited session organizers for helping to devise such an excellent scientific programme!

Like for previous ISNPS conferences, the *Journal of Nonparametric Statistics* will launch a special issue on the occasion of the *6th International Symposium on Nonparametric Statistics*. The special issue aims to publish on a fast track high quality papers on cutting-edge methods in nonparametric and semiparametric statistics presented at the conference. The review process will follow the general reviewing principles of *Journal of Nonparametric Statistics*, and submissions are possible from June 1 to September 30, 2024. More information is available at https://think.taylorandfrancis.com/special_issues/isnps2024/ All contributors to ISNPS 2024 are strongly encouraged to enhance the visibility and success of the conference by submitting their papers to the special issue.

We are extremely happy to welcome you in Braga and wish you a fruitful and enjoyable meeting!

Inês Sousa and Jacobo de Uña-Álvarez

ISNPS Steering Committee

- Patrice Bertail
- Ricardo Cao
- Hira Koul
- Jens-Peter Kreiss
- Soumendra Lahiri
- Michele La Rocca
- George Michailidis (Coordinator)
- Ursula U. Müller
- Efstathios Paparoditis
- Marianna Pensky
- Dimitris Politis
- Jeffrey Racine
- Ingrid Van Keilegom

Local Organizing Committee

- Ana P. Amorim (University of Minho)
- A. Manuela Gonçalves (University of Minho)
- Carla Moreira (University of Minho)
- Cecília Castro (University of Minho)
- Inês Sousa (University of Minho)
- Luís Machado (University of Minho)
- Marta Ferreira (University of Minho)
- Raquel Menezes (University of Minho)
- Susana Faria (University of Minho)
- Teresa Malheiro (University of Minho)

Scientific Committee

- Inês Sousa (co-chair) (University of Minho)
- Jacobo de Uña-Álvarez (co-chair) (University of Vigo)
- Harrison Zhou (Yale University)
- Igor Pruenster (Bocconi University)
- Ingrid Van Keilegom (KULeuven)
- Jeffrey Racine (Mc Master University)
- Luís Machado (University of Minho)
- Malka Gorfine (Tel Aviv University)
- Raquel Menezes (University of Minho)
- Sophie Dabo-Niang (University of Lille)
- Stathis Paparoditis (University of Cyprus)

ISNPS 2024

PROGRAMME

	Tue 25	Wed 26	Thu 27	Fri 28	Sat 29
09:00-10:00	Registration (9:00) Opening (9:30)	Contributed 2	Invited 5	Contributed 4	Invited 9
10:00-11:00	Peter Hall's Lecture Irène Gijbels	Keynote Talk 1 Peter Bühlmann		Keynote Talk 3 Peter Mueller	
11:00-11:30	Coffee-break	Coffee-break	Coffee-break	Coffee-break	Coffee-break
11:30-12:30	Contributed 1	Special Invited Talk 1 Andrew Barron	Keynote Talk 2 Jane-Ling Wang	Special Invited Talk 2 Wenceslao González-Manteiga	Keynote Talk 4 Silvia Gonçalves
12:30-13:30	Lunch-break	Lunch-break	Lunch-break	Lunch-break	Contributed 5
13:30-14:30	Invited 1	Invited 3	Invited 6	Invited 7	Lunch-break & farewell
14:30-15:30					
15:30-16:00	Coffee-break	Coffee-break	Coffee-break	Coffee-break	
16:00-17:00	Invited 2	ISNPS General Assembly	Contributed 3	Invited 8	
17:00-18:00		Invited 4	Excursions		
18:00-19:00	Welcome reception			Posters	
19:00 - 20:00					
20:00 - 22:00			Dinner		

Tuesday, 25 Jun

9:30 - 10:00 Opening
Room: Grande Auditorio

10:00 - 11:00 Peter Hall's Lecture
Chair: Ingrid Van Keilegom
Room: Grande Auditorio

10:00 Peter Hall and optimality and efficiency issues in smoothing in nonparametric and semiparametric estimation
Irène Gijbels

11:00 - 11:30 Coffee Break

11:30 - 12:30 Contributed 1

Bayesian nonparametrics
Chair: Inês Sousa
Room: Sala Polivalente 1.1

11:30 Bayesian semiparametric variable selection with shrinkage prior
Mingan Yang

11:50 Conic Sparsity: Estimation of Regression Parameters in Closed Convex Polyhedral Cones
Neha Agarwala, Anindya Roy, Arkaprava Roy

12:10 Stepwise Bayesian Optimization with Additive Kernels with High-Dimensional Constraints in Cost-Effectiveness Analysis
David Gomez-Guillén, Mireia Diaz, Jesus Cerquides

Dimension reduction techniques
Chair: James Allison
Room: Sala Polivalente 1.3

11:30 Covariate-informed reconstruction of functional data with missing fragments
Maximilian Ofner, Siegfried Hörmann

11:50 Extrinsic PCA
Vic Patrangenaru, Robert Paige, Ka Chun Wong, Mihaela Pricop-Jeckstadt

12:10 Multi-response Linear Regression Estimation Based on Low-rank Pre-smoothing
Xinle Tian

Functional data analysis 1
Chair: Andrea Meilán-Vila
Room: Grande Auditorio

11:30 Change point localisation and inference in fragmented functional data
Gengyu Xue, Haotian Xu, Yi Yu

11:50 Variable selection in nonparametric regression with functional and mixed covariates
Leonie Selk

12:10 Bias reduction for nonparametric estimation with functional data
Melanie Birke, Tim Greger

Regularized regression
Chair: Philippe Lambert
Room: Sala Polivalente 1.2

11:30 Density regression via Dirichlet process mixtures of normal structured additive regression models
Maria Xose Rodriguez Alvarez, Vanda Inácio, Nadja Klein

11:50 Cost-sensitive semi-parametric classification
Jorge C. Rella, Ricardo Cao, Juan M. Vilar Fernández

12:10 A non asymptotic analysis of the first component PLS regression
Luca Castelli, Clément Marteau, Irène Gannaz

Survival Analysis 1
Chair: Vlad Stefan Barbu
Room: Pequeno Auditorio

11:30 Conditional dependence structure under selection bias and informative censoring
Yassir Rabhi

11:50 Single-index model under (left-)truncation
Ewa Strzalkowska-Kominiak, Anna Herud

12:10 Local differential privacy in survival analysis using private failure indicators
Mikael Escobar-Bach, Maxime Egea

Time series 1
Chair: Angelina Roche
Room: Sala Polivalente 1.4

11:30 Local Whittle Estimation in Time-Varying Long Memory Series
Josu Arteche

11:50 A Test of Independence over Periods of Time for Locally Stationary Processes
Carina Beering

12:10 Adaptive prediction for functional time series
Hassan Maissoro, Valentin Patilea, Myriam Vimond

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 1

Nonparametric spatial statistics
Organizer: Rubén Fernández-Casal
Chair: Raquel Menezes
Room: Grande Auditorio

13:30 Testing a parametric circular regression function with spatially correlated data
Andrea Meilán-Vila, Mario Francisco-Fernández, Rosa M. Crujeiras

14:00 Spatiotemporal statistical analysis and inference techniques in Oceanography and Marine Science
Isabel Fuentes Santos

14:30 Approximating the cross-covariance of multivariate spatial processes through the direct covariances
Raquel Menezes, Pilar Garcia-Soidán

15:00 Nonparametric Geostatistics
Rubén Fernández-Casal

Adaptive functional data analysis
Organizer: Valentin Patilea
Chair: Valentin Patilea
Room: Pequeno Auditorio

13:30 A global test for heteroscedastic one-way FMANOVA with applications
Tianming Zhu, Jin-Ting Zhang, *Ming-Yen Cheng*

14:00 Minimax rates in regression models for functional data
Angelina Roche

14:30 Adaptive fPCA and score inference
Sunny Wang, Valentin Patilea

15:00 Locally Adaptive Online Functional Data Analysis
Valentin Patilea, *Jeffrey Racine*

Nonparametric methods in genetics and neuroscience

Organizer: Jere Koskela
Chair: Jere Koskela
Room: Sala Polivalente 1.2

13:30 Estimating multiple merger coalescents' characteristic measure
Arno Siri-Jégousse

14:00 Heavy-Tailed NGG-Mixture Models
Karla Vianey Palacios Ramirez

14:30 Asymptotic guarantees for Bayesian phylogenetic tree reconstruction
Alisa Kirichenko, Luke Kelly, *Jere Koskela*

15:00 Variable Selection through Penalized Regression: a stable approach
Ana Helena Tavares, Vera Afreixo, Gabriela Moura

Model specification and goodness-of-fit problems

Organizer: Juan Carlos Pardo-Fernández
Chair: Juan Carlos Pardo-Fernández
Room: Sala Polivalente 1.3

13:30 Two density-based tests for the k-sample problem with left-truncated data
Adrián Lago, Juan Carlos Pardo-Fernández, Jacobo de Uña-Álvarez, Ingrid Van Keilegom

14:00 Testing normality for many populations
M. Dolores Jiménez-Gamero

14:30 Testing for independence in vector autoregressive models
James Allison, Simos Meintanis, Joseph Ngatchou-Wandji

15:00 Tests of exogeneity in proportional hazards models with censored data
Ingrid Van Keilegom, Gilles Crommen, Jean-Pierre Florens

Assumption lean and other nonparametrics for health data

Organizer: Ronghui Xu
Chair: Ronghui Xu
Room: Sala Polivalente 1.4

13:30 Stage-Aware Learning for Dynamic Treatments
Hanwen Ye, Wenzhuo Zhou, Ruoqing Zhu, Annie Qu

14:00 Doubly Robust Estimation under Possibly Misspecified Marginal Structural Cox Model
Denise Rava, Ronghui Xu, Jelena Bradic, Jiyu Luo

14:30 Learning conditional average treatment effects using instrumental variables
Stijn Vansteelandt, Karla Diaz-Ordaz, Stephen O'Neill, Richard Grieve

15:00 Personalized reinforcement learning for healthcare: With applications to sepsis management in ICU
Linda Zhao, Junhui Cai, Ran Chen

Shape constrained statistical inference

Chair: Geurt Jongbloed
Organizer: Geurt Jongbloed
Room: Sala Polivalente 1.1

13:30 Doubly robust estimation and inference for a log-concave counterfactual density
Charles Doss

14:00 Semiparametric density estimation using copulas with log-concave marginals
Hanna Jankowski, Sawitree Boonpatcharanon

14:30 Density estimation using Total variation regularization
Arlene Kyoung Hee Kim, Adityanand Guntuboyina, Dohyeong Ki

15:00 Stereological determination of particle size distributions for similar convex bodies
Thomas van der Jagt, Geurt Jongbloed, Martina Vittorietti

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited 2

Network analysis and cluster analysis
Organizer:Anderson Ye Zhang
Chair: Anderson Ye Zhang
Room: Grande Auditorio

16:00 Interpretable network-assisted prediction
Tiffany Tang, *Elizaveta Levina*, Ji Zhu

16:30 Consistent community recovery from temporal and higher-order network interactions
Lasse Leskelä, Konstantin Avrachenkov, Maximilien Dreveton

17:00 Improved Mean Estimation in the Hidden Markovian Gaussian Mixture Model
Mohamed Ndaoud

Nonparametric methods to take advantage of auxiliary data in health settings
Organizer: Layla Parast
Chair: Layla Parast
Room: Pequeno Auditorio

16:00 Doubly Flexible Estimation under Label Shift
Yanyuan Ma

16:30 Conditional independence testing by comparing empirical conditional cumulative distribution functions
Boris Hejblum, Marine Gauthier, Sara Fallet, Rodolphe Thiébaut, Denis Agniel

17:00 A rank-based approach to evaluate a surrogate marker in a small sample setting
Layla Parast, Tianxi Cai, Lu Tian

17:30 Semiparametrically correcting for data quality issues to estimate whole-hospital, whole-body health from the EHR
Sarah Lotspeich, Joseph Rigdon

Random partitions and Bayesian dependent clustering
Organizer: Beatrice Franzolini
Chair: Beatrice Franzolini
Room: Sala Polivalente 1.1

16:00 Understanding partially exchangeable nonparametric priors for discrete structures
Beatrice Franzolini, Antonio Lijoi, Igor Pruenster, *Giovanni Rebaudo*

16:30 Informed Random Partition Models with Temporal Dependence
Garritt Page, Sally Paganin, Fernando Quintana

17:00 Continuous Clustering Models -- High-Dimensional Clustering Made Easy
Leo Duan, Arkaprava Roy

17:30 Bayesian nonparametric net survival estimation with clustering
Alan Riva-Palacio

Bayesian and mixed model approaches to optimal P-spline modelling
Organizer: Paul Eilers
Chair: Paul Eilers
Room: Sala Polivalente 1.2

16:00 A very short introduction to optimal smoothing with P-splines
Paul Eilers

16:30 Sparse mixed model P-splines with applications to multidimensional smoothing
Martin Boer

17:00 Fast Bayesian inference in complex additive models for censored data using Laplace P-splines
Philippe Lambert

17:30 Statistical modeling of infectious diseases with Laplacian-P-splines
Oswaldo Gressani

Statistics for non-stationary processes

Organizer: Patrice Bertail
Chair: Patrice Bertail
Room: Sala Polivalente 1.3

16:00 Models for Science Data with Hidden Periodic Structure
Antonio Napolitano

16:30 Harris recurrent Markov chains and nonlinear monotone cointegrated models
Carlos Fernández, Patrice Bertail, Cecile Durot

17:00 Optimal choice of bootstrap block length for periodically correlated time series
Anna Dudek, Patrice Bertail

17:30 Locally Stationary Spatial Processes
Soumendra Lahiri

Nonparametric methods for complex data

Organizer: Byeong Park
Chair: Byeong Park
Room: Sala Polivalente 1.5

16:00 Inference for Changing Periodicity, Smooth Trend and Covariate Effects in Time Series
Ming-Yen Cheng, David Siegmund, Shouxia Wang, *Lucy Xia*

16:30 Accelerated age-period-cohort models
Maria Dolores Martinez-Miranda, M. Luz Gamiz, Enno Mammen, Jens Perch Nielsen

17:00 A pseudo-metric between probability distributions based on depth-trimmed regions
Guillaume Staerman, *Pavlo Mozharovskyi*, Pierre Colombo, Stephan Clemencon, Florence D'Alche-Buc

17:30 Analysis in spectral domain for spatial data under fixed domain asymptotics
Chae Young Lim, Joonho Shin, Wei-Ying Wu

Advances in random networks

Organizer: Marianna Pensky
Chair: Elizaveta Levina
Room: Sala Polivalente 1.4

16:00 Signed Diverse Multiplex Networks: Clustering and Inference
Marianna Pensky

16:30 Random line graphs and edge-attributed network inference
Avanti Athreya, Zachary Lubberts, Youngser Park, Carey Priebe

17:00 Joint Spectral Clustering in Multilayer Degree-Corrected Stochastic Blockmodels
Zachary Lubberts, Joshua Agterberg, Jesus Arroyo

17:30 Intensity Profile Projection: A Framework for Continuous-Time Representation Learning for Dynamic Networks
Alexander Modell

18:00 - 19:00 Welcome Reception

Wednesday, 26 Jun

9:00 - 10:00 Contributed 2

Count data
Chair: Daniel Nevo
Room: Sala Polivalente 1.1

9:00 The minimax risk in nonparametric testing of discrete distributions for uniformity under missing ball alternatives
Alon Kipnis

9:20 Semiparametric test for overdispersed count data
Stefano Bonnini, Michela Borghesi

9:40 Multiple change-point detection in a Poisson process
Emilie Lebarbier

Extremes
Chair: Armelle Guillou
Room: Sala Polivalente 1.3

9:00 Extremal behaviour and convergence rates for sample-based geometric quantiles and half space depths
Marie Kratz

9:20 Spatio-temporal model for the occurrence of extreme events and inference on their extent
Ana C. Cebrian

9:40 Risk Assessment using a Semi-Parametric Approach
Ayana Mateus, Frederico Caeiro

Functional data analysis 2
Chair: Annika Betken
Room: Grande Auditorio

9:00 Measuring dependence between a scalar response and a functional covariate
Daniel Strenger, Siegfried Hörmann

9:20 Directional regularity: Achieving faster rates of convergence in multivariate functional data
Sunny Wang, *Omar Kassi*

9:40 Functional relevance based on continuous Shapley value
Pedro Delicado, Cristian Pachón-García

Smoothing methods
Chair: José E. Chacón
Room: Sala Polivalente 1.2

9:00 Bayesian wavelet regression using nonlocal prior mixtures and novel parameterizations
Nilotpal Sanyal

9:20 The Effective Degrees of Freedom in Kernel Density Estimation
Alex Trindade, Sofia Guglielmini, Igor Volobouev

9:40 Statistical modelling of the firing activity of grid cells using local polynomial kernel smoothing methods
Rida Ayyaz, Ioannis Papastathopoulos, Mathew Nolan

Survival Analysis 2
Chair: Rebecca Betensky
Room: Pequeno Auditorio

9:00 Increasing odds ratio, testing and applications
Paulo Eduardo Oliveira, Idir Arab, Tommaso Lando

9:20 Quantile modelling under dependent censoring
Myrthe D'Haen, Ingrid Van Keilegom, *Anneleen Verhasselt*

9:40 Testing for sufficient follow-up in survival data with a cure fraction
Tsz Pang Yuen, Eni Musta

Time series 2
Chair: Sophie Dabo
Room: Sala Polivalente 1.4

9:00 Bootstrap inference for group factor models
Benoit Perron, Silvia Gonçalves, Julia Koh

9:20 Distribution-free prediction intervals from interval score optimized pairs of nonparametric regression quantiles
Harry Haupt, *Joachim Schnurbus*, Ida Bauer

9:40 Pointwise spectral density estimation under local differential privacy
Karolina Klockmann, Tatyana Krivobokova, Cristina Butucea

10:00 - 11:00 Keynote Talk 1
Chair: Igor Pruenster
Room: Grande Auditorio

10:00 Causality-inspired Statistical Machine Learning
Peter Bühlmann

11:00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk 1
Chair: Sophie Dabo
Room: Grande Auditorio

11:30 Provably Fast and Accurate Estimation of Neural Nets
Andrew R. Barron

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 3

Nonstationary processes: theory and applications
Organizer: Anna Dudek
Chair: Anna Dudek
Room: Sala Polivalente 1.3

13:30 Bayesian nonparametric spectral analysis of locally stationary processes
Yifu Tang, *Claudia Kirch*, Jeong Eun Lee, Renate Meyer

14:00 Spectral analysis and subsampling for spectrally correlated processes
Anna E. Dudek, *Bartosz Majewski*

14:30 Trend estimation in a class of explosive count time series
Anne Leucht, Michael Neumann

15:00 Nonparametric hypothesis testing for the structure of spectrum of nonstationary processes
Jean-Marc Freyermuth

Theory and methods in Bayesian nonparametrics: recent advances
Organizer: Antonio Lijoi
Chair: Igor Pruenster
Room: Sala Polivalente 1.4

13:30 Functional connectivity across the human subcortical auditory system using an autoregressive matrix-Gaussian copula graphical model approach with partial correlations
Noirrit Kiran Chandra, Kevin Sitek, Bharath Chandrasekaran, Abhra Sarkar

14:00 Constrained Dirichlet Processes and Moment Condition Models
Jaeyong Lee

14:30 Distances on random probability measures
Marta Catalano

15:00 Bayesian Nonparametrics with the Martingale Posterior
Edwin Fong

Estimation and testing problems with survival data
Organizer: Jacobo de Uña-Álvarez
Chair: Jacobo de Uña-Álvarez
Room: Sala Polivalente 1.5

13:30 Surviving the multiple testing problem: RMST-based tests in general factorial designs
Merle Munko, Marc Ditzhaus, Dennis Dobler, Jon Genuneit

14:00 Bivariate dependent censoring with covariates
Noël Veraverbeke

14:30 A fully parametric model for non-proportional hazards survival analysis
María del Carmen Pardo, María del Mar Fenoy, Narayanaswamy Balakrishnan

15:00 Estimation and regression for sequentially-truncated data
Rebecca Betensky, Jing Qian, Erik Parner, Morten Overgaard

Large scale semi-parametric inference
Organizer: Omiros Papaspiliopoulos
Chair: Omiros Papaspiliopoulos
Room: Sala Polivalente 1.2

13:30 Partially factorized variational inference for high-dimensional mixed models
Max Goplerud, Omiros Papaspiliopoulos, Giacomo Zanella

14:00 Penalized likelihood estimation and inference in high-dimensional logistic regression
Ioannis Kosmidis, Philipp Sterzinger

14:30 On the role of parametrization in models with a misspecified nuisance component
Heather Battey

15:00 Empirical partially Bayes multiple testing and compound chi-square decisions
Nikolaos Ignatiadis, Bodhisattva Sen

Recent advances in time series and functional data analysis
Organizer: Alexander Aue
Chair: Jens-Peter Kreiss and Efstathios Paparoditis
Room: Sala Polivalente 1.1

13:30 Intrinsic and Extrinsic Graphical Models for Functional Data
Victor Panaretos

14:00 Integrative analysis of Riemannian and high-dimensional data
Eardi Lila, James Buenfil

14:30 A statistical framework for analyzing shape in a time series of random geometric objects
Anne van Delft, Andrew J. Blumberg

15:00 Prediction of Singular VARs and an Application to Generalized Dynamic Factor Models
Siegfried Hörmann, Gilles Nisol

Nonparametric causal inference

Organizer: Mats Stensrud
Chair: Mats Stensrud
Room: Pequeno Auditorio

13:30 A bipartite ranking approach to two-sample nonparametric hypothesis testing
Myrto Limnios, Stephan Clemencon, Nicolas Vayatis

14:00 A nonparametric Gail-Simon test and estimand for qualitative effect heterogeneity
Mats Stensrud, Aaron Hudson, Riccardo Brioschi, *Oliver Dukes*

14:30 Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space
Aaron Hudson

15:00 Kernel Debiased Plug-in Estimation
Ivana Malenica

Recent advances in semiparametric and nonparametric econometrics

Organizer: Juan Carlos Escanciano
Chair: Juan Carlos Escanciano
Room: Grande Auditorio

13:30 Regular identification of the ATE without the strict overlap condition
Telmo Pérez-Izquierdo

14:00 On the Existence and Information of Orthogonal Moments
Juan Carlos Escanciano, Facundo Argañaraz

14:30 Estimation and inference of panel data models with a generalized factor structure
Juan Manuel Rodriguez-Poo, Alexandra Soberon, Stefan Sperlich

15:00 Chi-square goodness-of-fit tests to check for conditional moment restrictions
Miguel A. Delgado, Antonio Raiola

15:30 - 16:00 Coffee Break

16:00 - 17:00 ISNPS General Assembly
Room: Grande Auditorio

17:00 - 19:00 Invited 4

Modern advances at the interface of statistical learning and inference

Organizer: Pragya Sur
Chair: Subhabrata Sen
Room: Grande Auditorio

17:00 Minimax estimation in Efron's two-groups model
Chao Gao

17:30 Denoising over network with application to partially observed epidemics
Olga Klopp, Claire Donnat, Nicolas Verzelen

18:00 Fast Linear Model Trees by PILOT
Peter Rousseeuw, Jakob Raymaekers, Tim Verdonck, Ruicong Yao

18:30 Adaptive Inference in Sequential Experiments
Cun-Hui Zhang, Mufang Ying, Koulik Khamaru

Causal inference for studying vaccine effects

Organizer: Daniel Nevo
Chair: Aaron Hudson
Room: Pequeno Auditorio

17:00 Evaluating immune correlates of protection in vaccine efficacy trials with stochastic-interventional causal effects
Nima Hejazi

17:30 Nonparametric Identification of Immunologic and Behavioral Effects in Vaccination Studies
Daniel Nevo, Mats Stensrud, Uri Obolski

18:00 Waning of treatment effects
Mats Stensrud, Matias Janvin

18:30 Vaccine effectiveness estimation under the test-negative design: identifiability and efficiency theory for causal inference under conditional and control exchangeability
Cong Jiang, Denis Talbot, Sara Carazo, *Mireille Schnitzer*

Computer-intensive methods for complex data

Organizer: Dimitris Politis
Chair: Dimitris Politis
Room: Sala Polivalente 1.1

17:00 Comparing many functional means
Stanislav Volgushev, Dehan Kong, Colin Decker

17:30 Statistical inference with optimal sampling
Alan Welsh, *Nan Zou*

18:00 Bootstrapping "Likehihood Ratio tests" under mispecification
Pascal Lavergne, *Patrice Bertail*

18:30 Bootstrap-assisted inference for weakly stationary time series
Yunyi Zhang

Topics in Econometrics: Big Data, Panel Estimation, and Forecasted Treatment

Organizer: Jeffrey Racine
Chair: Jeffrey Racine
Room: Sala Polivalente 1.2

17:00 A Robust Method for Microforecasting and Estimation of Random Effects
Silvia Sarpietro, Raffaella Giacomini, Sokbae Lee

17:30 Partial identification in nonlinear panels
Chris Muris

18:00 Fast Inference for Quantile Regression with Tens of Millions of Observations
Youngki Shin, Yuan Liao, Myung Hwan Seo, Sokbae Lee

18:30 Forecasted Treatment Effects
Irene Botosaru

Statistics for spatial and network data

Organizer: Soumendra Lahiri
Chairs: Soumendra Lahiri
Room: Sala Polivalente 1.3

17:00 Variance Estimation of Spectral Statistics for Spatial Processes using Subsampling
Souvick Bera, Daniel Nordman, *Soutir Bandyopadhyay*

17:30 Empirical likelihood inference in the frequency domain for dependent data
Dan Nordman, Haihan Yu, Mark Kaiser

18:00 Graph wavelet variances
Debashis Mondal, Rodney Fonseca, Aluisio Pinheiro

18:30 Conformal Prediction for Network-Assisted Regression
Robert Lunde, Elizaveta Levina, Ji Zhu

Topics in nonparametric and semiparametric econometrics

Organizers: Hira Koul and Indeewara Perera
Chair: Indeewara Perera
Room: Sala Polivalente 1.4

17:00 Estimation of Grouped Time-Varying Network Vector Autoregressive Models
Degui Li, Bin Peng, Songqiao Tang, Weibiao Wu

17:30 Inference of Unknown Semiparametric Transformation via Distribution Regression Estimation
Yi He, Juan-Juan Cai

18:00 Bootstrap specification tests for multivariate GARCH processes
Indeewara Perera, Kanchana Nadarajah

18:30 Regression Modelling under General Heterogeneity
Liudas Giraitis, Yufei Li, George Kapetanios

Semi- and non-parametric approaches for inference on high dimensional data
Organizer: Wen Zhou
Chair: Wen Zhou
Room: Sala Polivalente 1.5

17:00 Innovative unsupervised approach for simultaneous subgroup recovery and group-specific feature identification
Lyuou Zhang, Xiwei Tang, *Wen Zhou*

17:30 Multidimensional Signal-to-Noise Ratio Estimation for High Dimensional Random Effects Models under Heteroscedasticity
Xiaodong Li, Xiaohan Hu, Zhentao Li

18:00 Local perspectives in latent space network models
Lijia Wang, YX Rachel Wang, Xin Tong, Xiao Han, Yanhui Wu

18:30 Sparse Heteroskedastic PCA in High Dimensions
Zhao Ren, Rui Kang, Peiliang Zhang

Thursday, 27 Jun

9:00 - 11:00 Invited 5

Nonparametric estimation in high dimensions
Organizer: Zhou Fan
Chair: Nikolaos Ignatiadis
Room: Grande Auditorio

9:00 Rate Optimality and Phase Transition for User-Level Local Differential Privacy
Yi Yu

9:30 Fundamental limits of community detection from multi-view data
Subhabrata Sen, Xiaodong Yang, Buyu Lin

10:00 Spectrum-Aware Debiasing: A Modern Inference Paradigm with Applications to Principal Component Regression
Pragya Sur

10:30 Tightness of SDP and Burer-Monteiro Factorization for Phase Synchronization in High Noise Regime
Anderson Ye Zhang

Advanced inference of complex data
Organizer: Regina Liu
Chair: Dimitris Politis
Room: Pequeno Auditorio

9:00 Fair conformal prediction and risk control
Linjun Zhang

9:30 Shape analysis of functional data
Karthik Bharath

10:00 Selective inference with randomized Group LASSO estimators for general models
Snigdha Panigrahi

10:30 Nonparametric density estimation from streaming data
Aurore Delaigle

Causal inference in medical and public health studies
Organizer: Li Hsum
Chair: Yu Shen
Room: Sala Polivalente 1.1

9:00 Estimation of the complier causal hazard ratio under dependent censoring
Gilles Crommen, Jad Beyhum, Ingrid Van Keilegom

9:30 A Multi-State Modeling for the Cost-Effectiveness Analysis in Disease Prevention
Li Hsu

10:00 Using Joint Models for Longitudinal and Time-to-Event Data to Investigate the Causal Effect of Salvage Therapy after Prostatectomy
Jeremy Taylor, Dimitris Rizopoulos

10:30 Causal and Statistical Uncertainty for Individual-Level Causal Inference
Uri Shalit

Statistics for AI
Organizer: Yongdai Kim
Chair: Yongdai Kim
Room: Sala Polivalente 1.2

9:00 Minimax optimal density estimation using a shallow generative model
Chae Minwoo, Hyeok Kyu Kwon

9:30 A statistical analysis of an image classification problem
Sophie Langer, *Juntong Chen*, Johannes Schmidt-Hieber

10:00 Statistical Analysis on In-Context Learning
Masaaki Imaizumi

10:30 Optimal high-dimensional nonparametric regression with variational neural networks
Ilsang Ohn

Statistics for dependent data

Organizers: Jens-Peter Kreiss and Efstathios Paparoditis
Chairs: Jens-Peter Kreiss and Efstathios Paparoditis
Room: Sala Polivalente 1.3

9:00 A log-linear model for non-stationary time series of counts
Michael H. Neumann, Anne Leucht

9:30 Autoregressive Network: Sparsity and Degree Heterogeneity
Yutong Wang

10:00 Learning Graphical Models for nonstationary multivariate time series
Suhasini Subba Rao, Jonas Krampe

10:30 Asymptotic Theory for Constant Step Size Stochastic Gradient Descent
Jiaqi Li, Wei Biao Wu, Zhipeng Lou, *Stefan Richter*

Object Oriented Data Analysis: Trees and Graphs

Organizer: Steve Marron
Chair: Stephan Huckemann
Room: Sala Polivalente 1.4

9:00 Barycentric Subspace Analysis for Sets of Unlabelled Graphs
Anna Calissano, Elodie Maignant, Xavier Pennec

9:30 Sticky Flavors
Stephan F. Huckemann, Lars Lammers, Do Tran Van

10:00 Brownian motion, bridges and Bayesian inference in BHV tree space
Tom Nye, *William Woodman*

10:30 Estimating a mean tree for phylogenetic trees with missing taxa
Maryam Garba, Tom Nye

Analysis of curves

Organizer: Aurore Delaigle
Chair: Aurore Delaigle
Room: Sala Polivalente 1.5

9:00 Multivariate higher-order kernels
José E. Chacón, Tarn Duong

9:30 Robust estimation under small measurement errors
Michael Stewart, Alan Welsh

10:00 Partially observed functional data over non-Euclidean domains
Alessandro Palummo, Marco Stefanucci, Eleonora Arnone, *Laura M. Sangalli*

10:30 Kernel estimation for continuous-time semi-Markov processes
Vlad Stefan Barbu

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk 2

Chair: Malka Gorfine
Room: Grande Auditorio

11:30 Deep Learning for Censored Survival Data
Jane-Ling Wang

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 6

Current topics in biostatistics - nonparametric approaches

Organizer: Somnath Datta
Chair: Michael Daniels
Room: Grande Auditorio

13:30 Analysis of spatially clustered survival data with unobserved covariates using SBART
Debajyoti Sinha, Durbadal Ghosh, Antonio Linero, George Rust

14:00 A Bayesian nonparametric approach for nonignorable missingness in EHR data
Michael Daniels, David Lindberg, Sebastien Haneuse

14:30 Bayesian Nonparametric Modeling of Restricted Mean Survival Time: Subject Specific Inference and Average Treatment Effect
Sanjib Basu, Ruizhe Chen

15:00 Error Controlled Feature Selection for Ultrahigh Dimensional and Highly Correlated Feature Space Using Deep Learning
Taps Maiti

Regularized nonparametric regression for spatial and functional data

Organizer: Laura M. Sangalli
Chair: Eleonora Arnone
Room: Pequeno Auditorio

13:30 Sparsistency of estimators in semiparametric mixture of regression models
Abbas Khalili

14:00 A flexible framework for spatial quantile regression via PDE regularization
Cristian Castiglione

14:30 A regularized compositional functional concurrent regression model to investigate the dynamic relationship between causes of death and human longevity
Emanuele Giovanni Depaoli, *Marco Stefanucci*, Stefano Mazzuco

15:00 Function Estimation on Complex 3D Surfaces
Michelle Carey, Thiago Da Silva Cardoso

Recent advances in spatiotemporal data

Organizer: George Michailidis
Chair: George Michailidis
Room: Sala Polivalente 1.1

13:30 Semi-Parametric Inference for Doubly Stochastic Spatial Point Processes: An Approximate Penalized Poisson Likelihood Approach
Ali Shojaie, Si Cheng, Jon Wakefield

14:00 Likelihood Free Learning of Saptiotemporal Hawkes Processes
Moulinath Banerjee

14:30 Impulse Response Estimation in Large-scale Time Series
Sumanta Basu

15:00 Clustering of spatiotemporal processes using spectral analysis and applications
Soudeep Deb, *Sayar Karmakar*

Bayesian nonparametrics for complex data

Organizer: Ramses Mena
Chair: Ramses Mena
Room: Sala Polivalente 1.2

13:30 Finite population inference via martingales with a view towards quick counts
Carlos E. Rodríguez

14:00 Efficient estimation of the Posterior Similarity Matrix for Bayesian Nonparametric clustering
Johan van der Molen Moris

14:30 Clustering constrained on linear networks
Asael Fabian Martinez Martinez

15:00 Conditional partial exchangeability: a probabilistic framework for longitudinal and multi-view clustering
Beatrice Franzolini

Nonparametric statistics: methods and applications

Organizer: Sonali Das
Chair: Sonali Das
Room: Sala Polivalente 1.3

13:30 Co-variance Operator of Banach Valued Random Elements: U-Statistic Approach
Subhra Sankar Dhar

14:00 On A Goodness-of-fit Test for Elliptically Symmetric Distributions based on Scale-Scale Plots
Biman Chakraborty, Pritha Guha

14:30 Are winters getting shorter?
Anandamayee Majumdar, Sonali Das, Mehmet Balcilar, Levi Baguley, Siphumile Mangisa

15:00 Nonparametric Quantile Causality Assessment of Uncertainty and Gold: Multivariate and Bootstrap Extensions
Mehmet Balcilar, Rangan Gupta

Recent advances in depth and robust statistics

Organizer: Graciela Boente
Chair: Alicia Nieto-Reyes
Room: Sala Polivalente 1.4

13:30 Robust Functional Regression with Discretely Sampled Predictors
Ioannis Kalogridis

14:00 Local depth functions and clustering
Claudio Agostinelli, Giacomo Francisci, Anand Vidyashankar, Alicia Nieto-Reyes

14:30 Multivariate Singular Spectrum Analysis by Robust Diagonalwise Low-Rank Approximation
Mia Hubert, Fabio Centofanti, Peter Rousseeuw

15:00 On depth based two-sample tests: robustness in functional spaces
Alicia Nieto-Reyes, Felix Gnettner, Claudia Kirch

Cutting-edge machine learning for complex biomedical data

Organizer: Malka Gorfine
Chair: Malka Gorfine
Room: Sala Polivalente 1.5

13:30 Deep Learning of Partially Linear Cox Models: Error Rate and Selection Consistency
Yi Li

14:00 Post-Estimation Strategies in Sparse Semiparametric Models for High-Dimensional Data Application
S. Ejaz Ahmed

14:30 Accommodating Time-Varying Heterogeneity in Risk Estimation: A Transfer Learning Approach
Yu Shen, Jing Ning, Ziyi Li

15:00 Confidence Intervals and Simultaneous Confidence Bands Based on Deep Learning
Asaf Ben Arie, Malka Gorfine

15:30 - 16:00 Coffee Break

16:00 - 17:00 Contributed 3

Functional data analysis 3

Chair: Laura M. Sangalli
Room: Grande Auditorio

16:00 Functional approaches to nonparametric risk reserving using standard chain ladder data
Matus Maciak, Ivan Mizera, Michal Pesta

16:20 Statistical Inference For Spectral Means Of Hilbert Space Valued Random Processes
Daniel Rademacher, Jens-Peter Kreiss, Efstathios Paparoditis

16:40 Bayesian Variable Selection for Function-on-Scalar Regression Models: A Comparative Analysis
Camila de Souza, Pedro H. T. de Oliveira Sousa, Ronaldo Dias

Goodness-of-fit
Chair: M. Dolores Jiménez-Gamero
Room: Sala Polivalente 1.1

16:00 Tests of uniformity on the sphere with data-driven parameters
Alberto Fernández-de-Marcos, Eduardo García-Portugués

16:20 Addressing missing data challenges: A multivariate goodness-of-fit testing perspective
Danijel Aleksić, *Bojana Milošević*

16:40 Single-index quantile regression models: a new lack-of-fit test
Mercedes Conde-Amboage, Alvaro Arrojo-Vazquez

High-dimensional data 1
Chair: Claudio Agostinelli
Room: Sala Polivalente 1.2

16:00 Robustifying and simplifying high-dimensional regression with applications to yearly stock returns and telematics data
Michael Scholz, Maria Dolores Martinez-Miranda, Malvina Marchese, Jens Perch Nielsen

16:20 On the speed of the convergence of some kernel random forests.
Isidoros Iakovidis

16:40 Break detection procedures for high dimensional panel data
Charl Pretorius, Heinrich Roodt

Imperfectly observed data
Chair: Luís Machado
Room: Pequeno Auditorio

16:00 Infinitely divisible priors on exponent measures
Florian Brück

16:20 Tests of Missing Completely At Random based on sample covariance matrices
Alberto Bordino, Tom Berrett

16:40 Efficient quantile regression under censoring using Laguerre polynomials
Alexander Kreiss, Ingrid Van Keilegom

Network analysis
Chair: Zach Lubberts
Room: Sala Polivalente 1.3

16:00 Network evolution by clustering attachment
Natalia Markovich, Maksim Ryzhov, Marijus Vaiciulis

16:20 Equivariant and Invariant Modelling of Complex Data
Andreas Abildtrup Hansen, Aasa Feragen, Anna Calissano

16:40 Point processes on linear networks and how to address their comparison
Maria Isabel Borrajo García, Ignacio González-Pérez, Wenceslao González-Manteiga

Statistical learning and inference
Chair: Anna Calissano
Room: Sala Polivalente 1.4

16:00 The gROC curve and the optimal classification system
Pablo Martinez-Camblor, Sonia Perez Fernandez

16:20 Bayesian Analysis of a Multivariate Density Ratio Model
Victor De Oliveira

16:40 An inference method to deal with multiple causes of failure
Nora Villanueva, Marta Sestelo, Luís Meira-Machado, Javier Roca-Pardiñas

17:00 - 19:00 Excursions

20:00 - 22:00 Dinner

Friday, 28 Jun

9:00 - 10:00 Contributed 4

High-dimensional data 2
Chair: Natalie Neumeyer
Room: Sala Polivalente 1.2

9:00 Self-normalized sums in high dimension: which covariance estimator?
Emmanuelle Gautherat, Patrice Bertail, El Mehdi Issouani

9:20 Adaptive clustering through composite entropy Minimization
Thierry Dumont

9:40 Tail Inference with Probability Weighted Moments
Frederico Caeiro, Ayana Mateus, Dora Gomes

Multiple testing and simultaneous inference
Chair: Ruth Heller
Room: Sala Polivalente 1.1

9:00 Robust semi-parametric testing in generalized linear models with many responses
Jesse Hemerik

9:20 Post-hoc and Anytime Valid Permutation and Group Invariance Testing
Nick Koning

9:40 Blending Point-wise Inference and Cluster Mass Tests for powerful Massively Univariate Tests to control FWER in EEG data
Olivier Renaud, Jaromil Frossard

Nonparametric econometrics 1
Chair: Miguel A. Delgado
Room: Grande Auditorio

9:00 M-Estimation in Censored Regression Model using Instrumental Variables under Endogeneity
Swati Shukla

9:20 Weak convergence of the function-indexed sequential empirical process for nonstationary time series
Florian Scholze

9:40 Multiscale Comparison of Nonparametric Trend Curves
Marina Khismatullina, Michael Vogt

Nonparametric inference and estimation 1
Chair: Richard Samworth
Room: Pequeno Auditorio

9:00 Testing Conditional Dependence
Laura Freijeiro González, Wenceslao González-Manteiga, Manuel Febrero Bande

9:20 Inference for bivariate data with unobserved order
Laura Dumitrescu

9:40 Exponential bounds for penalized Hotelling statistics
El Mehdi Issouani, Patrice Bertail, Emmanuelle Gautherat

Set estimation and inference

Chair: Clément Levrard
Room: Sala Polivalente 1.3

9:00 Data Depth for Probability Measures
Pierre Lafaye de Micheaux, Pavlo Mozharovskyi, *Myriam Vimond*

9:20 Shape constraints beyond convexity
Alejandro Cholaquidis, Leonardo Moreno, *Beatriz Pateiro-López*

9:40 On Improved Semi-Parametric Bounds For Tail Probability And Expected Loss
Artem Prokhorov, Erick Li

10:00 - 11:00 Keynote Talk 3

Chair: Inês Sousa
Room: Grande Auditorio

10:00 Common atoms mixture models in two biostatistical inference problems
Peter Mueller

11:00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk 2

Chair: Jacobo de Uña-Álvarez
Room: Grande Auditorio

11:30 Specification tests for statistical models with recent results and applications
Wenceslao González-Manteiga

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 7

Statistics in the AI era: different perspectives

Organizer: Sophie Langer
Chair: Sophie Langer
Room: Grande Auditorio

13:30 Class probability matching for label shift adaptation
Annika Betken, Hongwei Wen, Hanyuan Hang

14:00 Differentially private penalized M-estimation via noisy optimization
Marco Avella Medina

14:30 Wasserstein Generative Adversarial Networks are Minimax Optimal Distribution Estimators
Arthur Stéphanovitch, Eddie Aamari, Clément Levrard

15:00 Dropout Regularization Versus L2-Penalization in the Linear Model
Gabriel Clara, Sophie Langer, Johannes Schmidt-Hieber

Bayesian nonparametrics for high-dimensional and complex models

Organizer: Ismaël Castillo
Chair: Ismaël Castillo
Room: Pequeno Auditorio

13:30 A variational Bayes approach to debiased inference in high-dimensional linear regression
Luke Travis, Ismaël Castillo, Alice L'Huillier, Kolyan Ray

14:00 Bayes in the extreme
Surya Tokdar

14:30 Almost-parallel Bayesian Gaussian Graphical Modelling in High-Dimensions
Deborah Sulem, David Rossell, Jack Jewson

15:00 Convergence rates of deep Gaussian process regression
Aretha Teckentrup

Conformal and simultaneous inference

Organizer: Etienne Roquain
Chair: Etienne Roquain
Room: Sala Polivalente 1.1

13:30 Combining exchangeable p-values
Matteo Gasparin, Aaditya Ramdas

14:00 Polya trees for nonparametric shrinkage estimation in high dimensional GLMs
Asaf Weinstein, Jonas Wallin, Daniel Yekutieli, Malgorzata Bogdan

14:30 Selecting informative conformal prediction sets with false coverage rate control
Ruth Heller, Etienne Roquain, Ulysse Gazin, Ariane Marandon

15:00 Transductive conformal inference with adaptive scores
Ulysse Gazin, *Gilles Blanchard*, Etienne Roquain

Extrapolation methods for extreme values

Organizer: Abdelaati Daouia
Chair: Abdelaati Daouia
Room: Sala Polivalente 1.2

13:30 Estimation of marginal excess moments for Weibull-type distributions
Yuri Goegebeur, Armelle Guillou, Jing Qin

14:00 A conditional tail expectation type risk measure for time series
Armelle Guillou, Yuri Goegebeur, Jing Qin

14:30 Functional Extreme-PLS
Stephane Girard, *Cambyse Pakzad*

15:00 Extreme expectile estimation for short-tailed data
Abdelaati Daouia, Simone Padoan, Gilles Stupfler

Recent advances in cure models

Organizer: Ricardo Cao
Chair: Ricardo Cao
Room: Sala Polivalente 1.3

13:30 A presmoothed estimator for the cure rate in mixture cure models
Ana López-Cheda, María Amalia Jácome Pumar, Samuel Saavedra

14:00 High dimensional mixture cure models: an application in cardio-oncology
Beatriz Piñeiro Lamas, Ricardo Cao, Ana López Cheda

14:30 Nonparametric inference for the mixture cure model with partially known cured observations
Wende Clarence Safari, Ignacio López-de-Ullibarri, *María Amalia Jácome Pumar*

15:00 Effect of a covariate in the cure rate of a mixture cure model using distance correlation
Blanca Estela Monroy Castillo, María Amalia Jácome Pumar, Ricardo Cao

High-dimensional regression

Organizer: Ursula Mueller
Chair: Ursula U. Müller
Room: Sala Polivalente 1.4

13:30 A Mean Field Approach to Empirical Bayes Estimation in High-dimensional Linear Regression
Bodhisattva Sen, Sumit Mukherjee, Subhabrata Sen

14:00 Statistical inference for the error distribution in functional linear models
Natalie Neumeyer

14:30 Adaptive variable selection in sparse nonparametric models
Natalia Stepanova, Marie Turcicova

15:00 Variable selection by voting
Ursula U. Müller

Recent advances in multivariate time series analysis

Organizer: Giovanni Motta
Chair: Giovanni Motta
Room: Sala Polivalente 1.5

13:30 Measuring and Predicting Cyclical Turning Points, Gaps, and Drawdowns
Tommaso Proietti

14:00 Non-parametric estimation of Dynamic Factor Models in the frequency domain
Giovanni Motta, Michael Eichler

14:30 Random matrices and spectral clustering for modeling high-dimensional self-similar systems
Gustavo Didier

15:00 A semi-parametric approach for clustering high-dimensional, non-stationary, auto-correlated time series
Qiyuan Wang, Giovanni Motta

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited 8

Identification and inference in semi- and non-parametric econometric models

Organizer: Karim Chalak
Chair: Karim Chalak
Room: Grande Auditorio

16:00 Inference for Regression with Variables Generated from Unstructured Data
Timothy Christensen

16:30 Inference on High Dimensional Selective Labeling Models
Shakeeb Khan, Elie Tamer, Qingsong Yao

17:00 Higher Order Moments for Differential Measurement Error, with Application to Tobin's q and Corporate Investment
Karim Chalak, Daniel Kim

17:30 Doubly Robust Bayesian Difference-in-Differences Estimators
Christoph Breunig, Ruixuan Liu, Zhengfei Yu

Advancements in semiparametric and large-scale inference

Organizer: Olga Klopp
Chair: Olga Klopp
Room: Pequeno Auditorio

16:00 Sparse additive models with discrete optimization
Peter Radchenko

16:30 Optimal convex M-estimation via score matching
Oliver Feng, Min Xu, Yu-Chun Kao, *Richard Samworth*

17:00 Nonparametric Maximum Likelihood Estimation of Monotone Binary Regression Models under Weak Feature Impact
Dario Kieffer, *Angelika Rohde*

17:30 Dynamic Topic Model
Cristina Butucea, Nayel Bettache, Tracy Ke

Structured nonparametric models

Organizer: Maria Dolores Martinez-Miranda
Chair: Maria Dolores Martinez-Miranda
Room: Sala Polivalente 1.1

16:00 Robust and flexible model selection for local linear conditional survival function estimation
Dimitrios Bagkavos, Jens Perch Nielsen, Montserrat Guillen

16:30 A semiparametric infinite-dimensional approach for factor analysis and dymamical multiple regression on manifolds
María Dolores Ruiz-Medina

17:00 A Complete Framework for Model-Free Difference-in-Differences Estimation
Stefan Sperlich

17:30 Smooth backfitting for additive hazard rates
Munir Eberhardt Hiabu, Stephan Bischofberger, Enno Mammen, Jens Nielsen

Bayesian sparse learning in high-dimensional problems

Organizer: Surya Tokdar
Chair: Surya Tokdar
Room: Sala Polivalente 1.2

16:00 Bayesian Covariance Estimation for Multi-group Matrix-variate Data
Elizabeth Bersson

16:30 Deep horseshoe Gaussian processes
Ismaël Castillo, Thibault Randrianarisoa

17:00 Bayesian inference in high-dimensional mixed frequency regression
Kshitij Khare

17:30 Bayesian Variable Selection in High-dimensional Settings with Grouped Covariates
Minerva Mukhopadhyay

Statistical methods for geometric inference and set estimation

Organizer: Beatriz Pateiro-López
Chair: Beatriz Pateiro-López
Room: Sala Polivalente 1.3

16:00 Wasserstein convergence of persistence diagrams on generic manifolds
Vincent Divol

16:25 Statistical difficulty of support estimation and dimensionality reduction
Clément Levrard, Eddie Aamari, Catherine Aaron, Clément Berenfeld

16:50 Confidence Regions for Filamentary Structures
Wanli Qiao

17:15 Highest density region estimation for manifold data
Diego Bolón, Rosa M. Crujeiras, Alberto Rodríguez-Casal

17:40 Two sample testing for isometry of two manifolds
Wolfgang Polonik, Eunseong Bae

Advances in functional data analysis

Organizer: Michelle Carey
Chair: Michelle Carey
Room: Sala Polivalente 1.4

16:00 Regression in quotient metric spaces with a focus on elastic curves
Lisa Steyer, Almond Stöcker, *Sonja Greven*

16:30 Space-time regression with non-stationary PDE penalization for the analysis of mobile phone data
Eleonora Arnone, Mara Sabina Bernardi, Laura Maria Sangalli, Piercesare Secchi

17:00 Statistical Analysis of Collections of Networks
Catherine Higgins, Michelle Carey, Hulin Wu

17:30 Block testing in precision matrix for functional data analysis
Alessia Pini, Marie Morvan, Madison Giacofci, Valerie Monbet

Recent advances in non-and semiparametric models in survival analysis

Organizer: Ingrid Van Keilegom
Chair: Ingrid Van Keilegom
Room: Sala Polivalente 1.5

16:00 Survival Estimation with Time-Varying Covariates Using Neural Networks
Bingqing Hu, *Bin Nan*

16:30 Cumulative Incidence Function Estimation Using Population-Based Biobank Data
Malka Gorfine, David M. Zucker, Shoval Shoham

17:00 A competing risks analysis with cause-specific cure
Eni Musta, Tijn Jacobs, Marta Fiocco

17:30 Conditional C-index for survival data with a cure fraction
Bo Han, Ingrid Van Keilegom, *Juan Carlos Pardo-Fernandez*

18:00 - 19:00 Posters

Scalar-on-Shape Regression Models in Functional Data Analysis
Sayan Bhadra, Anuj Srivastava

Modal Regression with Missing Response Data
Tomás R. Cotos-Yáñez, Rosa M. Crujeiras, Ana Pérez-González

Goodness-of-fit tests for circular data based on a Parzen-Rosenblatt type estimator
Carlos Tenreiro

Estimating asymptotic independence on the lower tail
Marta Ferreira

Multivariate asymptotic test of pairwise independence for orientations with the same symmetry group
Iva Karafiátová, Jakub Staněk, Zbyněk Pawlas

Inference on Data with Both Multiplicative and Additive Measurement Error
Yuxiang Zong, Yinfu Liu, Yanyuan Ma, Ingrid Van Keilegom

Semiparametric Generative Invariance
Carlos García Meixide, David Ríos Insua

Nonparametric methods for the extremal index estimation
Dora Prata Gomes, Manuela Neves

Bayesian Additive Regression Trees in Complex Survey
Abhishek Mandal

Modeling multivariate spatial dependency with copulas: a novel approach
Manuel Úbeda-Flores

Local logistic regression for dimension reduction in binary classification
Touqeer Ahmad, François Portier, Gilles Stupfler

Approximate Bayesian computation for Arrhenius relationship accelerated life tests
Lizanne Raubenheimer, *Neill Smit*

Explainable Deep Learning: a methodology to train Generalized Additive Model with deep neural networks
Ines Ortega-Fernandez, Marta Sestelo

Functional Data Analysis for Predicting Landed Fish Abundance per unit effort (LPUE)
Manuel Oviedo-de la Fuente, Raquel Menezes, Alexandra A. Silva

Extrinsic Principal Component Analysis
Ka Chun Wong, Vic Patrangenaru

Clustering cumulative incidence functions with clustcurv R package.
Marta Sestelo, Luís Meira-Machado, Nora Villanueva, Javier Roca-Pardiñas

Testing the fit of discrete response models with covariates
Leonard Santana, Simos Meintanis, Marius Smuts, Joseph Ngatchou Wandjii

New classes of tests for the Weibull distribution using Stein's method in the presence of random right censoring
Elzanie Bothma, James Allison, Jaco Visagie

On a new class of tests for the Pareto distribution using Fourier methods
Lethani Ndwandwe, James Allison, Marius Smuts, Jaco Visagie

On classes of consistent tests for the Pareto distribution based on a characterization involving order statistics
Thobeka Nombebe, James Allison, Joseph Ngatchou--Wandji, Leonard Santana

Saturday, 29 Jun

9:00 - 11:00 Invited 9

Statistics for a wise use of machine learning

Organizer: Stefan Sperlich
Chair: Stefan Sperlich
Room: Grande Auditorio

9:00 Fairness in Machine Learning : how AI creates and amplifies bias in the data.
Jean-Michel Loubes

9:30 Statistical methods for high-throughput experimental data
Tatyana Krivobokova, Gianluca Finocchio, Boris Maryasin

10:00 Trends in Statistical Deep Learning
Johannes Lederer

10:30 The implicit bias phenomenon in deep learning
Holger Rauhut

Nonparametric smoothing and regression for correlated observations

Organizers: Didier A. Girard and Sana Louhichi
Chair: Sana Louhichi
Room: Pequeno Auditorio

9:00 Inference on volatility estimation with missing data: a functional data approach
Mohamed Chaouch, Abdelbasset Djeniah, Amina Angelika Bouchentouf

9:30 Non-parametric statistic to test the equality of the health concentration curve and the 45 degree line
Taoufik Bouezmarni, Abderrahim Taamouti, Mohamed Doukali, Meryem Taleb Bendiab

10:00 On kernel density estimation for dependent data on Riemannian manifolds without boundary
Anne Francoise Yao, Vincent Monsan, Djack Guy-Aude Kouadio

10:30 Hyperparameters selection problems in nonparametric trend estimation: from statistics to machine learning
Sana Louhichi

Advances in directional statistics

Organizer: Thomas Verdebout
Chair: Rosa M. Crujeiras
Room: Sala Polivalente 1.1

9:00 On dependence analysis for circular data
Rosa M. Crujeiras

9:30 Kernel density estimation on the polysphere
Eduardo García-Portugués, Andrea Meilán-Vila

10:00 Conditional density estimation for spherical data
María Alonso-Pena, Paula Saavedra-Nieves

10:30 A novel data-based smoothing parameter for circular kernel density estimation
Jose Ameijeiras-Alonso

Network models and optimal prediction

Organizer: Moulinath Banerjee
Chair: Moulinath Banerjee
Room: Sala Polivalente 1.2

9:00 UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation
Debarghya Mukherjee, Jiawei Ge, Jiaqing Fan

9:30 A VAE-based Framework for Learning Multi-Level Neural Granger-Causal Connectivity
George Michailidis

10:00 Regression discontinuity design with explained score
Moulinat Banerjee, Debarghya Mukherjee, *Ya'acov Ritov*

Advances in financial econometrics
Organizer: Genaro Sucarrat
Chair: Genaro Sucarrat
Room: Sala Polivalente 1.3

9:00 Testing the zero-process of intraday financial returns for non-stationary periodicity
Ovidijus Stauskas, *Genaro Sucarrat*

9:30 Detection of breaks in weak location time series models with quasi-Fisher scores
Christian Francq, Lorenzo Trapani, Jean-Michel Zakoian

10:00 Finite moments testing in a general class of nonlinear time series models
Jean-Michel Zakoian, Christian Francq

10:30 Quantifying Uncertainty under Local Instability: a Dynamic Conformal approch to Electricity Price Forecasting
Alessandro Giovannelli, Tommaso Proietti, Andrea Cerasa, Fany Nan

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk 4
Chair: Jeffrey Racine
Room: Grande Auditorio

11:30 Bootstrapping out-of-sample predictability tests with real-time data
Silvia Gonçalves

12:30 - 13:30 Contributed 5

Biostatistics
Chair: M. Dolores Ruiz-Medina
Room: Sala Polivalente 1.1

12:30 Sample Size Planning for the Wilcoxon-Mann-Whitney Test with Dependent Replicates
Erin Sprünken, Frank Konietschke

12:50 Regression analysis for infectious disease modelling
Chengyuan Lu, Jelle Goeman, Hein Putter, Mar Rodriguez Girondo

13:10 Including time-dependent variables in ROC curve analyses
Arís Fanjul-Hevia, Juan Carlos Pardo-Fernandez, Wenceslao González-Manteiga

Nonparametric econometrics 2
Chair: Christian Francq
Room: Grande Auditorio

12:30 Revisiting Localized Technical Change: A Nonparametric Instrumental Regression Approach with Mixed-Type Endogenous Regressors
Davide Golinelli, Antonio Musolesi

12:50 One-step smoothing splines instrumental regression
Elia Lapenta, Jad Beyhum, Pascal Lavergne

13:10 The boosted Hodrick-Prescott filter, penalized least squares, and Bernstein polynomials
Keith Knight

Nonparametric inference and estimation 2
Chair: Eduardo García-Portugués
Room: Pequeno Auditorio

12:30 Evaluating Randomness Assumption: A Novel Graph Theoretic Approach for Linear and Circular Data
Shriya Gehlot, Arnab Kumar Laha

12:50 Optimal non-parametric estimation of distribution functions, convergence rates for Fourier inversion theorems and applications
Carlos Martins-Filho

13:10 Partial identification for a wide class of event time models in the presence of dependent censoring
Ilias Willems, Ingrid Van Keilegom, Jad Beyhum

13:30 - 14:30 Lunch & Farewell

ISNPS 2024

PROGRAMME WITH ABSTRACTS

Tuesday, 25 Jun

9:30 - 10:00 Opening
Room: Grande Auditorio

10:00 - 11:00 Peter Hall's Lecture
Chair: Ingrid Van Keilegom
Room: Grande Auditorio

10:00 **Peter Hall and optimality and efficiency issues in smoothing in nonparametric and semiparametric estimation**
Irène Gijbels

Abstract: Smoothing techniques for data on (subsets of) the real line are well developed. The development of smoothing techniques (such as kernel and local polynomial methods) for directional data has taken large steps forward in the last decades. Peter made seminal contributions in this area too, and research today builds further on these. Many challenging questions regarding statistical inference for directional data (and more generally data in non-Euclidean settings), need to get attention. Some selected topics will be discussed during this talk, ranging from kernel density estimation, to methods for high-dimensional data.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Contributed 1

Bayesian nonparametrics
Chair: Inês Sousa
Room: Sala Polivalente 1.1

11:30 **Bayesian semiparametric variable selection with shrinkage prior**
Mingan Yang

Abstract: Shrinkage priors have been widely used in linear regression models for variable selection. However, they are rarely used in mixed effects models. In addition, it is commonly assumed that the random effects have normal distributions. However, substantial deviation of normal distribution might potetnially impact the ultimate variable selection models. In this article, we address the problem of joint variable selection of both fixed and random effects in nonprametric models with shrinkage priors. An efficient Gibbs sampler is developed for posterior sampling. The approach is illustrated using a simulated exampe and a real data application.

11:50 **Conic Sparsity: Estimation of Regression Parameters in Closed Convex Polyhedral Cones**
Neha Agarwala, Anindya Roy, Arkaprava Roy

Abstract: Statistical problems often involve linear equality and inequality constraints on model parameters. Direct estimation of parameters restricted to general polyhedral cones, particularly when one is interested in estimating low dimensional features, may be challenging. We use a dual form parameterization to characterize parameter vectors restricted to lower dimensional faces of polyhedral cones and use the characterization to define a notion of 'sparsity' on such cones. We show that the proposed notion agrees with the usual notion of sparsity in the unrestricted case and prove the validity of the proposed definition as a measure of sparsity. The identifiable parameterization of the lower dimensional faces allows a generalization of popular spike-and-slab priors to a closed convex polyhedral cone. The prior measure utilizes the geometry of the cone by defining a Markov random field over the adjacency graph of the extreme rays of the cone. We describe an efficient way of computing the posterior of the parameters in the restricted case. We illustrate the usefulness of the proposed methodology for imposing linear equality and inequality constraints by using wearables data from the National Health and Nutrition Examination Survey (NHANES) actigraph study where the daily average activity profiles of participants exhibit patterns that seem to obey such constraints.

12:10 **Stepwise Bayesian Optimization with Additive Kernels with High-Dimensional Constraints in Cost-Effectiveness Analysis**
David Gomez-Guillén, Mireia Diaz, Jesus Cerquides

Abstract: Simulation models used in cost-effectiveness analysis serve as vital tools in healthcare decision-making processes, yet their accuracy heavily relies on the calibration of model parameters, often affected by uncertainty and numerous constraints. Bayesian optimization (BO) is a promising method to calibrate time-consuming models, aiming to reduce this uncertainty and align the simulation results with observed data. However, the high dimensionality of these models poses challenges for conventional BO approaches, requiring some modifications to ensure a computationally feasible calibration. We propose using BO techniques customized to calibrate cost-effectiveness models, focusing on addressing challenges posed by high dimensionality and particular structural patterns in these simulation models. Specifically, we explore the combined application of BO using Gaussian processes with additive kernels and a stepwise calibration approach, demonstrating their efficacy when used together. The integration of Gaussian processes with additive kernels allows the exploitation of the relationships within the model parameters, enhancing the search process. Moreover, the stepwise calibration approach proves particularly effective for simulation models with a sequential block structure, enabling efficient parameter search by breaking down a difficult optimization problem into several easier tasks. Additionally, we present a greedy method that takes advantage of the stepwise methodology to decompose some highly-dimensional constraints into a series of simple constraints. In particular, common constraints such as enforcing a set of ordered parameter values are considerably simplified, ensuring model coherence while maintaining computational efficiency. To illustrate the proposed techniques, a lung cancer simulation model with 99 parameters is employed as a case study. Through extensive experimentation with several methods, the efficacy of the customized BO approaches is demonstrated, showcasing significant improvements over traditional methods. Remarkably, our findings indicate that despite BO being typically recommended for optimizing time-consuming functions, it emerges as the fastest calibration method even for models with simulation times lower than a second.

Dimension reduction techniques
Chair: James Allison
Room: Sala Polivalente 1.3

11:30 **Covariate-informed reconstruction of functional data with missing fragments**
Maximilian Ofner, Siegfried Hörmann

Abstract: This talk addresses the reconstruction of partially observed functional data measured on a dense grid with additive noise. In this setting, we propose a novel procedure for recovering the missing fragments, using information from a completely observable subsample and potential covariates. Unlike traditional methods, our approach leverages factor models with increasing rank and avoids restrictive smoothness assumptions. We then discuss uniform convergence rates of our estimators under a triple asymptotic. Furthermore, we introduce a straightforward method for constructing simultaneous prediction bands. The methodology is finally illustrated by an application of real temperature curves.

11:50 **Extrinsic PCA**
Vic Patrangenaru, Robert Paige, Ka Chun Wong, Mihaela Pricop-Jeckstadt

Abstract: This is work on complex data, jointly with Robert L. Paige, Ka Chun Wond and Mihaela Pricop-Jeckstadt. The talk aims to develop a methodology for extrinsic principal component analysis on a manifold. Instead of using the local inverse of exponential map at the intrinsic mean of a random object on a manifold with a geodesic distance associated with a Riemannian structure (see Lin and Yao (2019)), that may lead to inadequate results as intrinsic PCA fails to take into account that often times geodesics fill densely a higher dimensional submanifold, one uses an embedding of the manifold into a numerical space, with the associated chord distance. The resulting extrinsic k-dimensonal principal subspace is obtained as preimage via the given embedding, of the intersection of the embdded manifold with the affine subspace spanned by the eigenvectors of the extrinsic covariance matrix corresponding to the k largest eigenvalues (see Patrangenaru and Ellingson(2015) for a definition), tied the at the embedded extrinsic population mean and by the orthocomplement of this tangent space in the ambient numerical space. Examples of estimation and data dimension reduction for analysis of random objects such as direct similarity shapes of contours or projective shapes of 3D landmark configurations extracted from digital camera images are also presented. Paige acknowledge NSF award 2311058, Patrangenaru acknowledge NSF award 2311059, and Pricop-Jeckstad acknowledge an M-Era Net award Horizons via European Research Council (ERC). [1] Patrangenaru, V. and Ellingson, L. E. (2015). Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis. CRC. [2] Zhenhua Lin, Fang Yao (2019). Intrinsic Riemannian functional data analysis. Ann. Statist. 47(6): 3533-3577

12:10 **Multi-response Linear Regression Estimation Based on Low-rank Pre-smoothing**
Xinle Tian

Abstract: Pre-smoothing is a technique aimed at increasing the signal-to-noise ratio in data to improve subsequent estimation and model selection in regression problems. However, pre-smoothing has thusfar been limited to the univariate response regression setting. Motivated by the widespread interest in multi-response regression analysis in many scientific applications, this article proposes a technique for data pre-smoothing in this setting based on low rank approximation. We establish theoretical results on the performance of the proposed methodology, and quantify its benefit empirically in a number of simulated experiments. We also demonstrate our proposed low rank pre-smoothing technique on real data arising from the environmental sciences.

Functional data analysis 1
Chair: Andrea Meilán-Vila
Room: Grande Auditorio

11:30 **Change point localisation and inference in fragmented functional data**
Gengyu Xue, Haotian Xu, Yi Yu

Abstract: We study the problem of change point localisation and inference for sequentially collected fragmented functional data, where each curve is observed only over discrete grids randomly sampled over a short fragment. The sequence of underlying covariance functions is assumed to be piecewise constant, with changes happening at unknown time points. To localise the change points, we propose a computationally efficient fragmented functional dynamic programming (FFDP) algorithm with consistent change point localisation rates. With an extra step of local refinement, we derive the limiting distributions for the refined change point estimators in two different regimes where the minimal jump size vanishes and where it remains constant as the sample size diverges. Such results are the first time seen in the fragmented functional data literature. As a byproduct of independent interest, we also present a non-asymptotic result on the estimation error of the covariance function estimator inspired by Lin et al. (2021). Our result accounts for the effects of the sampling grid size within each fragment under novel identifiability conditions. Extensive numerical studies are also provided to support our theoretical results.

11:50 **Variable selection in nonparametric regression with functional and mixed covariates**
Leonie Selk

Abstract: We consider a nonparametric regression model with multiple functional covariates, allowing for additional covariates of other types (categorical, continuous). The dependent variable can be categorical (binary or multi-class) or continuous, so that both classification and regression problems are considered. The estimation method is based on an extension of the Nadaraya- Watson estimator, where a kernel function is applied to a linear combination of distance measures, each computed on individual covariates. We are interested in distinguishing between relevant covariates and noise variables. It can be shown that a data-driven least squares cross-validation method can asymptotically remove irrelevant noise variables. Based on this understanding, a thresholded version of the extended Nadaraya-Watson estimator is proposed to perform variable selection.

12:10 **Bias reduction for nonparametric estimation with functional data**
Melanie Birke, Tim Greger

Abstract: Compared to nonparametric estimators in the multivariate setting, kernel estimators for functional data models have a larger order of bias, see e.g. Ferraty et al. (2007). This is problematic for constructing confidence regions or statistical tests since the bias might not be negligible. It stems from the fact that one sided kernels are used where already the first moment of the kernel is different from 0. This cannot be cured as in the multivariate setting by assuming the existence of higher derivatives and for all nonparametric functional kernel estimators the bias is of order h if h is the bandwidth used while the variance is of order 1/nF(h) where F(h) denotes the small ball probability. In this talk, we propose bias corrected estimators based on the idea in Cheng et al. (2018) which still have an easy structure but have a bias of order h^2 as in multivariate settings while the variance is of the same order as before. In addition we show asymptotic normality of such estimators and derive uniform rates. The performance of the estimator in finite samples is in addition checked in a simulation study. [1] M.-Y. Cheng, T. Huang, P. Liu, H. Peng (2018). Bias reduction for nonparametric and semiparametric regression models. Stat. Sin. 28, No. 4, Part 2, 2749--2770 [2] F. Ferraty, A. Mas, P. Vieu (2007). Nonparametric regression on functional data: inference and practical aspects. Aust. N. Z. J. Stat. 49, No. 3, 267--286

Regularized regression
Chair: Philippe Lambert
Room: Sala Polivalente 1.2

11:30 **Density regression via Dirichlet process mixtures of normal structured additive regression models**
Maria Xose Rodriguez Alvarez, Vanda Inácio, Nadja Klein

Abstract: Within Bayesian nonparametrics, dependent Dirichlet process mixture models provide a flexible approach for conducting inference about the conditional density function. However, several formulations of this class make either restrictive modelling assumptions or involve intricate algorithms for posterior inference. We present a flexible and computationally tractable model for density regression based on a single-weights dependent Dirichlet process mixture of normal distributions model for univariate continuous responses. We assume an additive structure for the mean of each mixture component and incorporate the effects of continuous covariates through smooth functions. The key components of our modelling approach are penalised B-splines and their bivariate tensor product extension. Our method also seamlessly accommodates categorical covariates, linear effects of continuous covariates, varying coefficient terms, and random effects, which is why we refer our model as a Dirichlet process mixture of normal structured additive regression models. A noteworthy feature of our method is its efficiency in posterior simulation through Gibbs sampling, as closed-form full conditional distributions for all model parameters are available. Results from a simulation study demonstrate that our approach successfully recovers the true conditional densities and other regression functionals in challenging scenarios. Applications to three real datasets further underpin the broad applicability of our method. An R package, DDPstar, implementing the proposed method is provided.

11:50 **Cost-sensitive semi-parametric classification**
Jorge C. Rella, Ricardo Cao, Juan M. Vilar Fernández

Abstract: Single index models (SIMs) are a class of semiparametric models in which a response variable is related to a weighted combination of explanatory variables via an unspecified function, on which any restriction is imposed. This gives SIMs interpretability and flexibility, facilitating the capture of complex data relationships. We extend SIMs to address the cost-sensitive classification problem by minimizing an instance-dependent loss function during model fitting. Leveraging the inherent flexibility of SIMs together with a cost-sensitive approach results in a powerful modelling approach. This is demonstrated through an extensive simulation study and the analysis of three real-world datasets, where the proposed methodology outperforms both previous cost-sensitive, parametric and semi-parametric approaches.

12:10 **A non asymptotic analysis of the first component PLS regression**
Luca Castelli, Clément Marteau, Irène Gannaz

Abstract: Partial Least Squares (PLS) regression is a dimension reduction technique used to handle high dimensionality. This method projects the data onto a carefully chosen subspace, considering successive correlations with the explanatory variable in order to improve the prediction quality. We focus our attention on the single component case, that provides a useful framework to understand the underlying mechanism. Despite its apparent simplicity, this scenario presents numerous statistical challenges. Specifically, the non-linearity of the corresponding estimator demands careful attention. We provide a non-asymptotic upper bound on the quadratic loss in prediction with high probability in a high dimensional regression context. The bound is attained thanks to a preliminary test on the first PLS component. In a second time, we extend these results to the sparse partial least squares approach. In particular, we exhibit upper bounds similar to those obtained with the lasso algorithm, up to an additional restricted eigenvalue constraint on the design matrix.

Survival Analysis 1
Chair: Vlad Stefan Barbu
Room: Pequeno Auditorio

11:30 **Conditional dependence structure under selection bias and informative censoring**
Yassir Rabhi

Abstract: Selection bias is common in cross-sectional surveys and prevalent-cohort studies on disease durations. Under biased sampling subjects with longer disease durations have a greater chance of being observed. As a result, covariate values linked to the longer survivors are favored by the sampling mechanism. When the sampled durations are also subject to right censoring, the censoring is informative. Modelling dependence structure without adjusting for these issues leads to biased results. In this presentation, we consider conditional copulas for modelling conditional dependence when the collected data are subject to selection bias and account for both informative censoring and covariate bias that are naturally linked to biased sampling. We address nonparametric estimation of conditional bivariate df, conditional copula, and conditional Kendall/Spearman measures for right-censored size-biased data. The proposed estimator for conditional bivariate df is a Hadamard-differentiable operator of the Kaplan-Meier estimator and the conditional empirical process. Based on this estimator, we devise estimators for conditional copula. The limiting processes of the estimators are studied and established. In addition, we introduce estimators for conditional Kendall/Spearman measures and study their weak convergences. The proposed method is then applied to analyze a set of right-censored size-biased data on survival with AIDS.

11:50 **Single-index model under (left-)truncation**
Ewa Strzalkowska-Kominiak, Anna Herud

Abstract: The single-index model is a useful tool to incorporate a d-dimensional vector of covariates X into a regression model avoiding the so called "curse of dimensionality". By assuming that there exists a vector of parameters θ so that the response variable depends only on the projection θ'X, one avoids a multivariate regression. While right censoring has been extensively addressed by Strzalkowska-Kominiak and Cao (2013), the context of truncated data has not been as thoroughly studied in the literature. In my work, I focus on truncation from the left, where the lifetime of interest T is observed only if L≤T, although the extending the methodology to right-truncation is straightforward. Estimating the parameter in the Single-Index model involves a likelihood-based approach adapted to left-truncated data and hazard rate estimation. This approach can be viewed as a nonparametric extension of the Cox regression model (see Kalbfleisch and Lawless (1991)), offering increased flexibility. I will assess the theoretical properties of the model and perform an extensive simulation study. [1] Strzalkowska-Kominiak, E., Cao, R. Maximum likelihood estimation for conditional distribution single-index models under censoring. Journal of Multivariate Analysis, 114 (2013), pp. 74-98. [2] Kalbfleisch, J. D., Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. Statistica Sinica, 19-32.

12:10 **Local differential privacy in survival analysis using private failure indicators**
Mikael Escobar-Bach, Maxime Egea

Abstract: Censored data analysis is always a difficult challenge due to the incompleteness nature of the observations. In practice, the censoring mechanism imposes a stringent setup for the statisticians and requires a dedicated methodology to obtain consistent statistical tools. Besides, studies in survival analysis usually apply to sensitive data where privacy protection is of crucial importance. In health care research or medicine, the release of shared databases has particularly increased the demand in guidelines for sanitized data. In this talk, we consider survival estimation with censored data under setups that preserve individual privacy. We provide an alpha-locally differentially private mechanism on failure indicators and propose a non-parametric kernel estimator for the cumulative hazard function. Under mild conditions, we obtain lowers bounds on the minimax rates of convergence and show that our estimator is minimax optimal under well-chosen bandwidths. The method is illustrated with numerical results on synthetic data.

Time series 1
Chair: Angelina Roche
Room: Sala Polivalente 1.4

11:30 **Local Whittle Estimation in Time-Varying Long Memory Series**
Josu Arteche

Abstract: The memory parameter is usually assumed to be constant in traditional long memory time series. We relax this restriction by considering the memory a time varying function that depends on a finite number of parameters. A time-varying Local Whittle estimator of these parameters, and hence of the memory function, is proposed. Its consistency and asymptotic normality are shown for locally stationary and locally non-stationary long memory processes, where the spectral behaviour is restricted only at frequencies close to the origin. Its good finite sample performance is shown in a Monte Carlo exercise and in two empirical applications, highlighting its benefits over the fully parametric Whittle estimator proposed by Palma and Olea (2010). Standard inference techniques for the constancy of the memory are also proposed based on this estimator.

11:50 **A Test of Independence over Periods of Time for Locally Stationary Processes**
Carina Beering

Abstract: Considering two locally stationary processes, we can look for independence at a given point in time or, more thoroughly, over periods of time up to the whole observed time horizon. Thus, we developed the testing procedure in Beering (2021), which uses a characteristic function-based weighted distance inspired by the distance covariance defined by Székely et al. (2007) and its use by Jentsch et al.(2020), further to include the time aspect. The bootstrap-driven testing procedure is trialed by a simulation study aiming to detect independence as well as dependence of different forms. Lastly, the test is used to check for independence between different sensor outputs belonging to a bridge monitoring system as a real world application. [1] Beering, C. (2021). A functional central limit theorem and its bootstrap analogue for locally stationary processes with application to independence testing. Dissertation. Technische Universität Braunschweig. [2] Jentsch, C., Leucht, A., Meyer, M. and Beering, C. (2020). Empirical characteristic functions-based estimation and distance correlation for locally stationary processes. Journal of Time Series Analysis 41, 110-133. [3] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics 35, 2769-2794.

12:10 **Adaptive prediction for functional time series**
Hassan Maissoro, Valentin Patilea, Myriam Vimond

Abstract: An adaptive curve prediction method is proposed for a stationary functional time series (FTS). The sample paths of the functional time series are assumed to be irregular and observed with error at discrete points in the domain. Our approach is based on the best linear unbiased predictor (BLUP), which requires knowledge of several functions characterizing the FTS and the errors' conditional variance. We propose adaptive nonparametric estimators for the mean, covariance and autocovariance functions. The estimators adapt to the local regularity of the FTS, leading to a reduction in risk prediction. A simulation study and a real data application illustrate the good performance of the new predictor.

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 1

Nonparametric spatial statistics
Organizer: Rubén Fernández-Casal
Chair: Raquel Menezes
Room: Grande Auditorio

13:30 **Testing a parametric circular regression function with spatially correlated data**
Andrea Meilán-Vila, Mario Francisco-Fernández, Rosa M. Crujeiras

Abstract: In this work, new approaches for testing a parametric regression function for linear-circular regression models (circular response and Euclidean covariates) with spatially correlated errors are proposed and analyzed. The test statistics employed in these procedures are based on a comparison between a (non-smoothed or smoothed) parametric fit under the null hypothesis and a nonparametric estimator of the circular regression function. Notice that, in this framework, a suitable measure of circular distance must be employed. The null hypothesis that the regression function belongs to a certain parametric family is rejected if the distance between both fits exceeds a certain threshold. To perform the parametric estimation, procedures based on least squares or maximum likelihood are used. For the nonparametric alternative, a local linear-type estimator is considered. For practical application, different bootstrap methods are designed, and their performance is analyzed and compared in empirical experiments.

14:00 **Spatiotemporal statistical analysis and inference techniques in Oceanography and Marine Science**
Isabel Fuentes Santos

Abstract: The fast technological advances in areas such as remote sensing, which favor the intensification of monitoring systems, almost real time access to high resolution satellite data, and open data policies, have implied a substantial increase in data availability for oceanographic and marine research. In particular, scientist can obtain almost real time physical (temperature, salinity) and biogeochemical (chlorophyll, DOC, CO2...) satellite data with high spatiotemporal resolution which,, in combination with information provided by cruises and monitoring systems, are key to address issues such as local scale estimation of climate change effects in coastal areas. In the other hand, the use of acoustic telemetry for fish species monitoring is a useful tool for the management of fisheries and marine protected areas. This increasing data availability implies an important challenge for marine researchers, as the widely used classical statistical techniques do not allow a proper incorporation of the spatial and temporal components. This talk introduces some contributions of spatiotemporal statistics to marine science. We shall see the application of spatiotemporal nonparametric inference methods to the analysis of high resolution primary production satellite data in coastal areas, and the contribution to mussel culture planning. We also show the application of nonparametric point processes inference to research in fish population dynamics. Finally, we discuss some open challenges, such as the development of inference tools for a proper analysis of data with high spatial and temporal sparseness collected in coastal areas.

14:30 **Approximating the cross-covariance of multivariate spatial processes through the direct covariances**
Raquel Menezes, Pilar Garcia-Soidán

Abstract: The assumption of stationarity simplifies tackling inference problems for spatial data. However, an appropriate characterization of the correlation structure of the underlying process is usually required and this issue is particularly complex in the multivariate scenario. For instance, under second-order stationarity, the main difficulties are due to the number of covariance functions that must be estimated, as well as to the relationships among them, which convey that the characterization of these functions cannot be accomplished in an independent way. Different approaches have been suggested in the statistics literature to overcome the aforementioned drawbacks, although, to our knowledge, none of these proposals aims to solely involve the direct covariances to address the estimation of the whole dependence structure. This is the main goal of the current work, which is supported by the fact that the approximation of the direct covariances is quite simpler than that of the cross-covariances, particularly for heterotopic data. Taking this into account, we suggest estimating each cross-covariance through an appropriate linear combination of the direct covariances of the involved variables. Additionally, we intend to quantify the error committed in the estimation, when proceeding in this way. We conclude that the resulting error directly depends on the correlation degree between the variables involved. Numerical studies for simulated data and a real data set, obtained from environmental monitoring, are included to illustrate the application of the proposed methods. Acknowledgement: First author acknowledges the FCT Foundation for funding her research through projects with DOI 10.54499/UIDP/00013/2020 and DOI 10.54499/UIDB/00013/2020.

15:00 **Nonparametric Geostatistics**
Rubén Fernández-Casal

Abstract: The aim of this talk is to review some of the nonparametric geostatistical methods that the author has been working on in recent years, including some details about their implementation in practice with the R package npsp (https://rubenfcasal.github.io/npsp). We will focus on homoscedastic spatial processes with a deterministic trend and an isotropic variogram, although this methodology could be applied to heteroscedastic or spatiotemporal data. The first step is the nonparametric modeling of the process, consisting in the joint estimation of the trend and the variogram (np.fitgeo function). An iterative algorithm is used for this purpose, selecting the bandwidth for trend estimation by a method that takes the spatial dependence into account (h.cv function) and correcting for the bias due to the use of residuals in the estimation of the variogram (np.svariso.corr function). From these estimates, predictions at unobserved locations can be obtained by residual kriging (np.kriging function). A nonparametric bootstrap method has also been developed that allows additional inferences to be made about the process, such as obtaining confidence (or prediction) intervals, testing hypotheses or constructing unconditional risk maps. This bootstrap method can be modified to make inferences about the distribution of the response at new locations conditional on the observed values, for example, for the construction of conditional risk maps.

Adaptive functional data analysis
Organizer: Valentin Patilea
Chair: Valentin Patilea
Room: Pequeno Auditorio

13:30 **A global test for heteroscedastic one-way FMANOVA with applications**
Tianming Zhu, Jin-Ting Zhang, *Ming-Yen Cheng*

Abstract: Multivariate functional data are prevalent in various fields such as biology, climatology, and finance. Motivated by the World Health Data applications, in this study, we propose and examine a global test for assessing the equality of multiple mean functions in multivariate functional data. This test addresses the one-way Functional Multivariate Analysis of Variance (FMANOVA) problem, which is a fundamental issue in the analysis of multivariate functional data. While numerous analysis of variance tests have been proposed and studied for univariate functional data, only a limited number of methods have been developed for the one-way FMANOVA problem. Furthermore, our global test has the ability to handle heteroscedasticity in the unknown covariance function matrices that underlie the multivariate functional data, which is not possible with existing methods. We establish the asymptotic null distribution of the test statistic as a chi-squared-type mixture, which depends on the eigenvalues of the covariance function matrices. To approximate the null distribution, we introduce a Welch--Satterthwaite type chi-squared-approximation with consistent parameter estimation. The proposed test exhibits root-n consistency, meaning it possesses nontrivial power against a local alternative. Additionally, it offers superior computational efficiency compared to several permutation-based tests. Through simulation studies and applications to the World Health Data, we highlight the advantages of our global test.

14:00 **Minimax rates in regression models for functional data**
Angelina Roche

Abstract: In recent decades, significant research efforts have focused on regression models that involve functional data, which are data that can be modeled as samples of random functions. The minimax rates for the functional linear model and the fully nonparametric model are now well understood, although some aspects of these models still requires further exploration. However, for other models, like the single index model or the models with sparsity, the minimax rates are still unknown. The objective of this presentation is to provide a brief overview of the current state of knowledge regarding these models, as well as ongoing research on them.

14:30 **Adaptive fPCA and score inference**
Sunny Wang, Valentin Patilea

Abstract: Functional data analysis almost always involves smoothing discrete observations into curves, because they are never observed in continuous time and rarely without error. Although smoothing parameters affect the subsequent inference, data-driven methods for selecting these parameters are not very popular, frustrated by the difficulty of using all the information shared by curves while being computationally efficient. On the one hand, smoothing individual curves in an isolated, albeit sophisticated way, ignores useful signals present in other curves. On the other hand, bandwidth selection by automatic procedures such as cross-validation after pooling all the curves together quickly become computationally unfeasible due to the large number of data points. We present a new data-driven, adaptive kernel smoothing, specifically tailored for functional principal components analysis (fPCA) through the derivation of explicit risk bounds for the eigen-elements. Equipped with the principal components, we conduct inference for the scores by using effective linear integration rules inspired by recent Monte Carlo integration procedures. For our study, both common and independent design cases are allowed, with no Gaussianity assumption required.

15:00 **Locally Adaptive Online Functional Data Analysis**
Valentin Patilea, *Jeffrey Racine*

Abstract: One drawback with classical smoothing methods (kernels, splines, wavelets etc.) is their reliance on assuming the degree of smoothness (and thereby assuming continuous differentiability up to some order) for the underlying object being estimated. However, the underlying object may in fact be irregular (i.e., non-smooth and even perhaps nowhere differentiable) and, as well, the (ir)regularity of the underlying function may vary across its support. Elaborate adaptive methods for curve estimation have been proposed, however, their intrinsic complexity presents a formidable and perhaps even insurmountable barrier to their widespread adoption by practitioners. We contribute to the functional data literature by providing a pointwise MSE-optimal, data-driven, iterative plug-in estimator of “local regularity” and a computationally attractive, recursive, online updating method. In so doing we are able to separate measurement error “noise” from “irregularity” thanks to “replication”, a hallmark of functional data. Our results open the door for the construction of minimax optimal rates, “honest” confidence intervals, and the like, for various quantities of interest.

Nonparametric methods in genetics and neuroscience

Organizer: Jere Koskela
Chair: Jere Koskela
Room: Sala Polivalente 1.2

13:30 **Estimating multiple merger coalescents' characteristic measure**
Arno Siri-Jégousse

Abstract: In this talk, I will expose a technique to infer the multiple merger coalescent model from genetic data taken at one single time. Coalescents are random genealogical models that are now accepted by the community to represent the gene tree of a population. Their characteristic measure gives some information on populations evolutionary history such as skewed offspring, selection, bottlenecks... This technique goes in two steps: 1) estimating the coalescent rates from the genetic data and in particular the so-called weighted Site Frequency Spectrum, and 2) estimating the characteristic measure Lambda of the multiple merger coalescent thanks to a non-parametric estimation method based on Bernstein polynomials. This estimation is useful to infer the best model to represent the genealogical tree of a population, but also works well to reject commonly used coalescent models (such as Beta-coalescents, Kingman coalescent....). The error of consistency generated by this method can also be established. This is a joint work with Verónica Miró Pina and Émilien Joly

14:00 **Heavy-Tailed NGG-Mixture Models**
Karla Vianey Palacios Ramirez

Abstract: Heavy tails are often found in practice, and yet they are an Achilles heel of a variety of mainstream random probability measures such as the Dirichlet process (DP). The first contribution of this paper focuses on characterizing the tails of the so-called normalized generalized gamma (NGG) process. We show that the right tail of an NGG process is heavy-tailed provided that the centering distribution is itself heavy-tailed; the DP is the only member of the NGG class that fails to obey this convenient property. A second contribution of the paper rests on the development of two classes of heavy-tailed mixture models and the assessment of their relative merits. Multivariate extensions of the proposed heavy-tailed mixtures are devised here, along with a predictor-dependent version, to learn about the effect of covariates on a multivariate heavy-tailed response. The simulation study suggests that the proposed method performs well in various scenarios, and we showcase the application of the proposed methods in a neuroscience dataset.

14:30 **Asymptotic guarantees for Bayesian phylogenetic tree reconstruction**
Alisa Kirichenko, Luke Kelly, *Jere Koskela*

Abstract: Bayesian tree reconstruction procedures constitute a central class of algorithms for inferring shared ancestry among DNA sequence samples in phylogenetics. We derive tractable criteria for the consistency of these algorithms as the sequence length and number of sequences increase. Our results encompass several Bayesian algorithms in widespread use in software such as BEAST, MrBayes, and RevBayes. Unlike almost all existing asymptotic guarantees for tree reconstruction, we require no discretization or boundedness assumptions on branch lengths. Our flexible results are easy to adapt to variations of the underlying inference problem. We demonstrate the practicality of our criteria on two examples with binary trees: a Kingman coalescent prior on rooted, ultrametric trees, and an independence prior on unrooted, unconstrained trees. In both cases, our convergence rate matches known frequentist results, obtained under stronger boundedness assumptions, up to logarithmic factors. Our results also apply to non-binary tree models.

15:00 **Variable Selection through Penalized Regression: a stable approach**
Ana Helena Tavares, Vera Afreixo, Gabriela Moura

Abstract: In this work, we propose a stable and accurate procedure to perform feature selection in datasets with a much higher number of predictors than individuals, as in genome-wide association studies. Due to the instability of feature selection in the presence of many potential predictors, we suggest a variable selection procedure that combines the results of repeated applications of a penalized regression model. We use a weighted formulation to identify the most important variables and to define the corresponding regression coefficient using a measure of the relative quality of fit of each model. To illustrate the stable variable selection procedures, we apply them to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for the investigation of single nucleotide polymorphism (SNP) predictors associated with Alzheimer's disease. The dataset comprises 451 individuals and 518,257 SNPs. Our experiments show the potential of the new method for stable feature selection.

Model specification and goodness-of-fit problems

Organizer: Juan Carlos Pardo-Fernández
Chair: Juan Carlos Pardo-Fernández
Room: Sala Polivalente 1.3

13:30 **Two density-based tests for the k-sample problem with left-truncated data**
Adrián Lago, Juan Carlos Pardo-Fernández, Jacobo de Uña-Álvarez, Ingrid Van Keilegom

Abstract: The comparison of populations can be addressed in many different ways, depending on the interests of the researcher or the a priori information that one has about the target distribution. One can, for example, employ the well-known t-test or ANOVA test for to compare means under normality. If one is interested in any kind of differences between the populations, distinct functions related to a random variable can be employed. In this case, we will focus on tests based on estimators of the density function, which are known to be more powerful than the distribution-based tests in certain situations. On the other hand, there exist situations in which the observation of individuals is partially hidden by the presence of another random variable; this is called truncation. Literature referring to hypothesis contrasts under truncation is, until now, vaguely developed, being the log-rank and generalizations of it the most employed tests. From the estimation of the cumulative distribution function, one can define a proper estimator of the density function under one-sided truncation and study tests based on it. In this talk, two different tests based on such estimator will be proposed to address the k-sample problem with left-truncated data. Their asymptotic distributions will be studied and, due to the impossibility of its application in practice, two slightly different bootstrap resampling plans will be proposed to approximate the null distributions of the tests statistics. Both the validation of these methods and the choice of the smoothing parameter will be addressed via Monte Carlo simulations. Moreover, both tests will be compared to existing tests in the left-truncation framework, such as the Kolmogorov-Smirnov and the log-rank tests. Two real datasets regarding pregnancy and unemployment times will be employed to exemplify the performance of the proposed tests.

14:00 **Testing normality for many populations**
M. Dolores Jiménez-Gamero

Abstract: we study the problem of simultaneously testing that each of k independent samples come from a normal population. The means and variances of those populations may differ. The proposed procedures are based on the BHEP test and they allow k to increase, which can be even larger than the sample sizes.

14:30 **Testing for independence in vector autoregressive models**
James Allison, Simos Meintanis, Joseph Ngatchou-Wandji

Abstract: We consider tests for serial independence of arbitrary finite order for the innovations in vector autoregressive models. The tests are expressed as L2-type criteria involving the difference of the joint empirical characteristic function and the product of corresponding marginals. Asymptotic as well as Monte-Carlo results are presented.

15:00 **Tests of exogeneity in proportional hazards models with censored data**
Ingrid Van Keilegom, Gilles Crommen, Jean-Pierre Florens

Abstract: Consider a duration time T, a possibly endogenous covariate Z and a vector of exogenous covariates X such that T=phi(Z,X,U) is increasing in U with U ~ U[0,1]. Moreover, let T be right-censored by a censoring time C such that only their minimum, denoted by the follow-up time Y=min{T,C}, is observed. In this paper, we construct a test statistic for the hypothesis that Z is exogenous w.r.t. T, where T, given Z and X, is assumed to follow a proportional hazards model. Note that this is equivalent to testing whether U is independent of Z. Our test makes use of an instrumental variable W that is independent of U, since it can be shown that Z is exogenous w.r.t. T if and only if V_T = F_{T | Z,X}(T | Z,X) is independent of W. We prove some asymptotic properties of the proposed test, provide possible bootstrap approximations for the critical value and show that we have a good finite sample performance via simulations. Lastly, we give an empirical example using The National Job Training Partnership Act (JTPA) Study.

Assumption lean and other nonparametrics for health data
Organizer: Ronghui Xu
Chair: Ronghui Xu
Room: Sala Polivalente 1.4

13:30 **Stage-Aware Learning for Dynamic Treatments**
Hanwen Ye, Wenzhuo Zhou, Ruoqing Zhu, Annie Qu

Abstract: Recent advances in dynamic treatment regimes (DTRs) provide powerful optimal treatment searching algorithms, which are tailored to individuals' specific needs and able to maximize their expected clinical benefits. However, existing algorithms could suffer from insufficient sample size under optimal treatments, especially for chronic diseases involving long stages of decision-making. To address these challenges, we propose a novel individualized learning method which estimates the DTR with a focus on prioritizing alignment between the observed treatment trajectory and the one obtained by the optimal regime across decision stages. By relaxing the restriction that the observed trajectory must be fully aligned with the optimal treatments, our approach substantially improves the sample efficiency and stability of inverse probability weighted based methods. In particular, the proposed learning scheme builds a more general framework which includes the popular outcome weighted learning framework as a special case of ours. Moreover, we introduce the notion of stage importance scores along with an attention mechanism to explicitly account for heterogeneity among decision stages. We establish the theoretical properties of the proposed approach, including the Fisher consistency and finite-sample performance bound. Empirically, we evaluate the proposed method in extensive simulated environments and a real case study for COVID-19 pandemic.

14:00 **Doubly Robust Estimation under Possibly Misspecified Marginal Structural Cox Model**
Denise Rava, Ronghui Xu, Jelena Bradic, Jiyu Luo

Abstract: We address the challenges posed by non-proportional hazards and informative censoring, offering a path toward more meaningful causal inference conclusions. We start from the marginal structural Cox model, which has been widely used for analyzing observational studies with survival outcomes, and typically relies on the inverse probability weighting method. The latter hinges upon a propensity score model for the treatment assignment, and a censoring model which incorporates both the treatment and the covariates. In such settings, model misspecification can occur quite effortlessly, and the Cox regression model's non-collapsibility has historically posed challenges when striving to guard against model misspecification through augmentation. We introduce an augmented inverse probability weighted estimator which, enriched with doubly robust properties, paves the way for integrating machine learning and a plethora of nonparametric methods, effectively overcoming the challenges of non-collapsibility. The estimator extends naturally to estimating a time-average treatment effect when the proportional hazards assumption fails. We closely examine its theoretical and practical performance. Finally, its application to a dataset reveals insights into the impact of mid-life alcohol consumption on mortality in later life.

14:30 **Learning conditional average treatment effects using instrumental variables**
Stijn Vansteelandt, Karla Diaz-Ordaz, Stephen O'Neill, Richard Grieve

Abstract: Driven by clinical inquiries regarding the efficacy of emergency surgery for gastrointestinal conditions, I will discuss the estimation of conditional average treatment effects (CATE) in the presence of unmeasured confounding using instrumental variables. While data-adaptive methods (e.g. statistical learning or machine learning) offer relief from concerns related to model misspecification bias inherent in parametric approaches, they introduce their own challenges, notably regularization bias and potential overfitting. Specifically, slow convergence rates affecting the first stage (machine learning based) regression of exposure on instrument and covariates, may propagate into the CATE estimates, resulting in poor accuracy. Synthetic data simulations confirm this, but moreover reveal poor performance of existing strategies for constructing a so-called Neyman-orthogonal learner, which is nonetheless designed to mitigate regularization bias in first-stage predictions. To address this, I will propose an alternative Neyman-orthogonal learner by strategically tailoring first-stage predictions to excel in their ultimate task: delivering CATE estimates with low regularization bias. Simulation studies validate substantial enhancements in performance, underscoring the effectiveness of the proposed approach.

15:00 **Personalized reinforcement learning for healthcare: With applications to sepsis management in ICU**
Linda Zhao, Junhui Cai, Ran Chen

Abstract: In numerous fields such as healthcare, public policy, and e-commerce, a primary objective is to make multiple decisions simultaneously in a dynamic and personalized fashion. This sequential decision-making process is especially relevant in healthcare for developing personalized treatment plans. The main challenge stems from the dynamic and personalized nature of the process -- each patient's history and unique responses to treatments significantly influence their current and future care. To tackle these challenges, we develop a personalized reinforcement learning algorithm that provides optimal and interpretable personalized treatment decisions. Focusing on sepsis management in ICUs, a condition as the main cause of mortality in hospitals accounting for more than 20 billion of total costs yet no consensus on optimal treatment strategies, we demonstrate the value of our algorithm on the ICU data from five Boston hospitals. We show that our algorithm can outperform standard care by providing more effective and personalized treatment plans for sepsis patients, showcasing the potential of our approach to improve outcomes and reduce costs in complex healthcare settings.

Shape constrained statistical inference
Organizer: Geurt Jongbloed
Chair: Geurt Jongbloed
Room: Sala Polivalente 1.1

13:30 **Doubly robust estimation and inference for a log-concave counterfactual density**
Charles Doss

Abstract: We consider the problem of causal inference based on observational data (or the related missing data problem) with a binary or discrete treatment variable. In that context we study estimation and inference for the density of the unobserved counterfactual variable; the counterfactual density provides more nuanced information than the counterfactual mean (or average treatment effect) does. We impose the shape-constraint of log-concavity on the counterfactual density, and then develop doubly robust estimators of the log-concave counterfactual density (based on an augmented inverse-probability weighted pseudo-outcome). Log-concavity allows us to develop an estimator that does not depend on a bandwidth or tuning parameter (once we have the pseudo-outcomes). We show the consistency in various global metrics of that estimator, and also find the estimator's pointwise limit distribution. We also develop asymptotically valid pointwise confidence intervals for the counterfactual density (without needing to estimate the density's second derivative, a challenging task, which such confidence intervals sometimes require). We demonstrate our procedures' performances in simulations and on a real dataset.

14:00 **Semiparametric density estimation using copulas with log-concave marginals**
Hanna Jankowski, Sawitree Boonpatcharanon

Abstract: We propose a semi-parametric estimation approach for multivariate densities using copulas with log-concave marginals. The benefit of the approach is two-fold: we obtain dimension-free convergence rates and reduce computational complexity of the estimators. The drawback is of course the additional constraint imposed by the parametric copula. In this talk I will discuss the details of the approach and the rates of convergence obtained. The performance in the well- and misspecified settings is studied via simulations, and I will show the results here as well.

14:30 **Density estimation using Total variation regularization**
Arlene Kyoung Hee Kim, Adityanand Guntuboyina, Dohyeong Ki

Abstract: We study the problem of nonparametric estimation of an unknown density on the real line using penalized maximum likelihood, where the penalty is based on the total variation of an appropriate derivative of the log-density. This estimator has been around for a while, but many theoretical properties including rates of convergence are unavailable. We prove such rates of convergence, and explain connections to shape-constrained density estimation.

15:00 **Stereological determination of particle size distributions for similar convex bodies**
Thomas van der Jagt, Geurt Jongbloed, Martina Vittorietti

Abstract: In the classical Wicksell problem 3D spheres of random sizes are positioned in an opaque medium according to a homogeneous Poisson process. Taking a planar section of the medium, the problem is to determine the size distribution of the spheres using the observed circular section profiles. We study a generalization of this problem. Consider an opaque medium which contains 3D particles. All particles are convex bodies of the same shape, but they vary in size. The particles are randomly positioned and oriented within the medium and cannot be observed directly. Taking a planar section of the medium we obtain a sample of observed 2D section profile areas of the intersected particles. The distribution of interest is the underlying 3D particle size distribution, for which we obtain an identifiability result. Moreover, we propose a consistent likelihood-based estimator for this size distribution.

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited 2

Network analysis and cluster analysis
Organizer: Anderson Ye Zhang
Chair: Anderson Ye Zhang
Room: Grande Auditorio

16:00 **Interpretable network-assisted prediction**
Tiffany Tang, *Elizaveta Levina*, Ji Zhu

Abstract: Machine learning algorithms often assume that training samples are independent. When data points are connected by a network, it creates dependency between samples, which is a challenge, reducing effective sample size, and an opportunity to improve prediction by leveraging information from network neighbors. Multiple prediction methods taking advantage of this opportunity are now available. Many methods including graph neural networks are not easily interpretable, limiting their usefulness in the biomedical and social sciences, where understanding how a model makes its predictions is often more important than the prediction itself. Some are interpretable, for example, network-assisted linear regression, but generally do not achieve similar prediction accuracies as more flexible models. We bridge this gap by proposing a family of flexible network-assisted models built upon a generalization of random forests (RF+), which both achieves highly-competitive prediction accuracy and can be interpreted through feature importance measures. In particular, we provide a suite of novel interpretation tools that enable practitioners to not only identify important features that drive model predictions, but also quantify the importance of the network contribution to prediction. This suite of general tools broadens the scope and applicability of network-assisted machine learning for high-impact problems where interpretability and transparency are essential.

16:30 **Consistent community recovery from temporal and higher-order network interactions**
Lasse Leskelä, Konstantin Avrachenkov, Maximilien Dreveton

Abstract: Community recovery is the task of learning a latent community structure from interactions in a population of N nodes. Efficient algorithms for sparse binary pairwise interaction data are well known, and so are their consistency properties with respect to data sampled from the stochastic block model (SBM), the canonical model for random graphs with a community structure. Instead of a binary variable indicating whether or not an interaction occurs, we often also observe a category, value, or shape of an interaction. This motivates the definition of a generalised SBM in which interactions can be of arbitrary type, including categorical, numeric, and vector-valued, and not excluding even more general objects such as Markov chains or Poisson processes. For this model, I will discuss information-theoretic bounds which characterise the existence of consistent estimators in terms of data sparsity, statistical similarity between intra- and inter-block interaction distributions, and the shape and size of the interaction space. Temporal networks with time-correlated interaction patterns of length T provide an important model instance, for which consistency can be analysed with respect to either N or T, or both, approaching infinity. For temporal networks, I will discuss the statistical accuracy of spectral methods for community recovery. Time permitting, I will also highlight recent findings related to data sets involving higher-order interactions which can be modelled using hypergraph and tensor SBMs.

17:00 **Improved Mean Estimation in the Hidden Markovian Gaussian Mixture Model**
Mohamed Ndaoud

Abstract: In this talk we will investigate the problem of center estimation in the Gaussian Mixture Model with Hidden Markovian labels. We first study the limitations of existing results in the high dimensional setting and then propose a minimax optimal procedure for the problem of mean estimation. Along the way, we also propose some adaptive variants of our procedure and show that adaptive rates might suffer from the curse of dimensionality.

Nonparametric methods to take advantage of auxiliary data in health settings

Organizer: Layla Parast
Chair: Layla Parast
Room: Pequeno Auditorio

16:00 **Doubly Flexible Estimation under Label Shift**
Yanyuan Ma

Abstract: In studies ranging from clinical medicine to policy research, complete data are usually available from a population P, but the quantity of interest is often sought for a related but different population Q which only has partial data. We consider the setting when both outcome Y and covariate X are available from P but only X is available from Q, under the label shift assumption; i.e., the conditional distribution of X given Y is the same in the two populations. To estimate the parameter of interest in Q by leveraging information from P, three ingredients are essential: (a) the common conditional distribution of X given Y, (b) the regression model of Y given X in P, and (c) the density ratio of the outcome Y between the two populations. We propose an estimation procedure that only needs some standard nonparametric technique to approximate the conditional expectations with respect to (a), while by no means needs an estimate or model for (b) or (c); i.e., doubly flexible to the model misspecifications of both (b) and (c). This is conceptually different from the well-known doubly robust estimation in that, double robustness allows at most one model to be misspecified whereas our proposal can allow both (b) and (c) to be misspecified. This is of particular interest in label shift because estimating (c) is difficult, if not impossible, by virtue of the absence of the Y-data from Q. While estimating (b) is occasionally off-the-shelf, it may encounter issues related to the curse of dimensionality or computational challenges. We develop the large sample theory for the proposed estimator, and examine its finite-sample performance through simulation studies as well as an application to the MIMIC-III database.

16:30 **Conditional independence testing by comparing empirical conditional cumulative distribution functions**
Boris Hejblum, Marine Gauthier, Sara Fallet, Rodolphe Thiébaut, Denis Agniel

Abstract: We propose a novel, distribution-free, and flexible conditional independence test that rely on conditional cumulative distribution functions, estimated through repeated logistic regressions. We provide the asymptotic distribution of the this test statistic as well as a permutation test. This new method, called citcdf, tests the association of a quantitative variable with one or many variables of interest (that can be either continuous or discrete), while potentially adjusting for additional covariates. We also provide a grouped version of this test. We study this novel testing method in the context of Differential Expression Analysis (DEA) for single-cell RNA-seq data. State-of-the-art methods for single-cell RNA (scRNA-seq) sequencing scRNA-seq DEA often rely on strong distributional assumptions that are difficult to verify in practice. Furthermore, while the increasing complexity of clinical and biological single-cell studies calls for greater tool versatility, the majority of existing methods only tackle the comparison between two conditions at once. citcdf substantially expands the possibilities for scRNA-seq DEA studies to test such complex hypotheses: it demonstrates good statistical performance in various simulation scenarios considering complex experimental designs (i.e. beyond the two condition comparison), while retaining competitive performance with state-of-the-art methods in a two-condition benchmark. We apply citcdf to a large scRNA-seq dataset of 84,140 SARS-CoV-2 reactive CD8+ T cells, in order to identify the diffentially expressed genes across 3 groups of COVID-19 severity (mild, hospitalized, and ICU) while accounting for seven different cellular subpopulations.

17:00 **A rank-based approach to evaluate a surrogate marker in a small sample setting**
Layla Parast, Tianxi Cai, Lu Tian

Abstract: In clinical studies of chronic diseases, the effectiveness of an intervention is often assessed using “high cost” outcomes that require long-term patient follow-up and/or are invasive to obtain. While much progress has been made in the development of statistical methods to identify surrogate markers, that is, measurements that could replace such costly outcomes, they are generally not applicable to studies with a small sample size. These methods either rely on nonparametric smoothing which requires a relatively large sample size or rely on strict model assumptions that are unlikely to hold in practice and empirically difficult to verify with a small sample size. In this paper, we develop a novel rank-based nonparametric approach to evaluate a surrogate marker in a small sample size setting. The method developed in this paper is motivated by a small study of children with nonalcoholic fatty liver disease (NAFLD), a diagnosis for a range of liver conditions in individuals without significant history of alcohol intake. Specifically, we examine whether change in alanine aminotransferase (ALT; measured in blood) is a surrogate marker for change in NAFLD activity score (obtained by biopsy) in a trial, which compared Vitamin E (n = 50) versus placebo (n = 46) among children with NAFLD.

17:30 **Semiparametrically correcting for data quality issues to estimate whole-hospital, whole-body health from the EHR**
Sarah Lotspeich, Joseph Rigdon

Abstract: The allostatic load index (ALI) is an informative summary of whole-person health, drawing upon biomarkers to measure lifetime strain. Borrowing data from electronic health records (EHR) is a natural way to estimate whole-person health and identify at-risk patients on a large scale. However, these routinely collected data contain missingness and errors, and ignoring these data quality issues can lead to biased statistical results and incorrect clinical decisions. Validation of EHR data (e.g., through chart reviews) can provide better-quality data, but realistically, only a subset of patients' data can be validated. Thus, we consider strategic ways to harness the error-prone ALI from the EHR to target the most informative patient records for validation. Specifically, the validation study is designed to achieve the best statistical precision to quantify the association between ALI and healthcare utilization in a logistic regression model. Further, we propose a semiparametric maximum likelihood estimator for this model, which robustly corrects data quality issues in unvalidated records while preserving the power of the full cohort. Through simulations and an application to the EHR of an extensive academic learning health system, targeted partial validation and the semiparametric estimator are shown to be effective and efficient ways to correct data quality issues in EHR data before using them in research.

Random partitions and Bayesian dependent clustering
Organizer: Beatrice Franzolini
Chair: Beatrice Franzolini
Room: Sala Polivalente 1.1

16:00 **Understanding partially exchangeable nonparametric priors for discrete structures**
Beatrice Franzolini, Antonio Lijoi, Igor Pruenster, *Giovanni Rebaudo*

Abstract: Species sampling models provide a general framework for random discrete distributions that are tailored for exchangeable data. However, they fall short when used for modeling heterogeneous data collected from related sources or distinct experimental conditions. To address this, partial exchangeability serves as the ideal probabilistic framework. While numerous models exist for partially exchangeable observations, a unifying framework, like species sampling models, is currently missing for this framework. Thus, we introduce multivariate species sampling models, a general class of models characterized by their partially exchangeable partition probability function. They encompass existing nonparametric models for partial exchangeable data, highlighting their core distributional properties. Our results allow the study of the induced dependence structure and facilitate the development of new models. This is a joint work with Beatrice Franzolini, Antonio Lijoi, and Igor Pruenster.

16:30 **Informed Random Partition Models with Temporal Dependence**
Garritt Page, Sally Paganin, Fernando Quintana

Abstract: Model-based clustering is a powerful tool that is often used to discover hidden structure in data by grouping observational units that exhibit similar response values. Recently, clustering methods have been developed that permit incorporating an "initial" partition informed by expert opinion. Then, using some similarity criteria, partitions different from the initial one are down weighted, i.e. they are assigned reduced probabilities. These methods represent an exciting new direction of method development in clustering techniques. We add to this literature a method that very flexibly permits assigning varying levels of uncertainty to any subset of the partition. This is particularly useful in practice as there is rarely clear prior information with regards to the entire partition. Our approach is not based on partition penalties but considers individual allocation probabilities for each unit (e.g., locally weighted prior information). We illustrate the gains in prior specification flexibility via simulation studies and an application to a dataset concerning spatio-temporal evolution of PM_10 measurements in Germany.

17:00 **Continuous Clustering Models -- High-Dimensional Clustering Made Easy**
Leo Duan, Arkaprava Roy

Abstract: Model-based clustering is routinely used in statistics. Most existing models, such as Gaussian mixture, are based on countable mixture, where the latent mixing distribution is assumed to be discrete. While the countable mixture idea works well in low dimensions, it suffers from an explosion of computing and modeling complexity in high dimensions. In this talk, I will present a new class of "continuous clustering models" based on uncountable mixture, in which the mixing distribution is continuous but has group-wise concentrations. To induce such concentrations, we take a Bayesian approach and regularize the pairwise difference matrix with commonly seen continuous shrinkage priors, such as horseshoe. Like continuous shrinkage models in variable selection tasks, these continuous clustering models enjoy outstanding computing performance, have a low computing cost per MCMC iteration, and a rapid mixing of Markov chains. I will discuss a few advanced applications, including manifold clustering of images and hybrid mixture clustering of time series. I will demonstrate how continuous clustering removes the modeling and computing burdens in those applications.

17:30 **Bayesian nonparametric net survival estimation with clustering**
Alan Riva-Palacio

Abstract: The gold standard for net survival inference is the frequentist nonparametric Pohar-Perme estimator (Perme et. al 2012). In contrast with overall survival, net survival considers time-to-event observations within larger populations which are different; thus allowing for proper comparison of the survival experience across such populations. A usual example is when net survival is related to death of patients caused by a specific type of cancer independently of cause of death due to the population. We propose a Bayesian nonparametric alternative for net survival estimation based on neutral to the right priors (Doksum 1974) which is analytically and computationally tractable. In particular we provide a posterior characterization for the model and implement an algorithm for clustering of exact observations due to population or net survival.

Bayesian and mixed model approaches to optimal P-spline modelling

Organizer: Paul Eilers
Chair: Paul Eilers
Room: Sala Polivalente 1.2

16:00 **A very short introduction to optimal smoothing with P-splines**
Paul Eilers

Abstract: In the last few decades, P-splines have found a prominent place in semiparametric modeling. They combine a B-splines basis with a discrete penalty on their coefficients. Instead of trying to optimize the number of splines, a generous number is chosen, and the penalty is used to tune smoothness. In the last few years new developments have led to effective and very efficient algorithms, using mixed model or Bayesian ideas. I will give a short introduction to P-splines, setting the stage for the three other speakers in this session, who will describe the new developments and illustrate them with challenging applications.

16:30 **Sparse mixed model P-splines with applications to multidimensional smoothing**
Martin Boer

Abstract: Penalized regression using B-splines (P-splines; Eilers and Marx 1996) can be computationally efficient, because of the local character of B-splines. The corresponding linear equations are sparse and can be solved quickly. However, the main problem is to find the optimal values for the penalty parameters. A good way to approach this problem is to use mixed models and restricted maximum likelihood (REML; Patterson and Thompson, 1971). Several methods have been proposed to transform the original penalized B-spline model to a mixed model. A problem with most existing transformations to mixed models is that the local character of the B-splines is lost, which reduces the computational efficiency. In Boer (2023) a new method was proposed, using a sparse transformation to mixed models. This method is computationally more efficient than other approaches. For example, to model the spatial variation of a precipitation data set in the USA, the new approach is more than 100 times faster than the original implementation by Rodriguez-Alvarez et al. (2014). The new method is better scalable and therefore can be applied to large datasets. Two examples will be presented using the R-package LMMsolver (Boer 2023) on CRAN. The first example is using three-dimensional P-splines to model hyperspectral imaging. The second example is estimating crop yield using a spatial-temporal P-splines model for satellite data. [1] Boer, M.P. (2023) Tensor product P-splines using a sparse mixed model formulation. Statistical Modelling 23 465–479. [2] Eilers, P.H.C., Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. Stat. Sci. 11, 89–121 [3] Patterson, H.D. and Thompson R. (1971) Recovery of inter-block information when block sizes are unequal. Biometrika, 58, 545—554 [4] Rodriguez-Alvarez, M.X. et al. (2014) Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. Statistics and Computing, 25, 941–957.

17:00 Fast Bayesian inference in complex additive models for censored data using Laplace P-splines

Philippe Lambert

Abstract: Laplace P-spline models (LPS) combine the P-spline smoother with Laplace approximations to perform fast Bayesian inference without the need for Markov chain Monte Carlo (MCMC) methods. Based on analytical approximations to the posterior of the penalty parameters, the amount of smoothing can be selected automatically using iterative methods. The uncertainty in the estimation of the model components such as additive terms in complex regression model can be quantified by accounting for the uncertainty in the selection of the penalty parameters. It relies on closed form approximations to the marginal posterior of the regression and spline parameters. The procedure is extremely fast and provides estimators with excellent frequentist properties. The methodology will be illustrated on double additive models for censored data, including semi-parametric location-scale and cure survival models with time-varying covariates. [1] Lambert, P. and Kreyenfeld, M. (2023). Exogenous time-varying covariates in double additive cure survival model with application to fertility. arXiv:2302.00331. [2] Lambert, P. and Gressani, O. (2023). Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models. Statistical Modelling, 23(5-6): 409–423. [3] Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data. Computational Statistics and Data Analysis, 161: 107250. [4] Gressani,O. and Lambert, P. (2021). Laplace approximation for fast Bayesian inference in generalized additive models based on P-splines. Computational Statistics and Data Analysis, 154: 107088. [5] Marx, D.M. and Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. Computational Statistics and Data Analysis, 28(2): 193-209.

17:30 Statistical modeling of infectious diseases with Laplacian-P-splines

Oswaldo Gressani

Abstract: The recent SARS-CoV-2 pandemic has underlined the crucial role of statistical modeling as it forms the core backbone to compute estimates of key epidemiologic quantities from infectious disease data. Having statistical methods that deliver state-of-the art analytical tools is not only important for understanding the transmission dynamics of a pathogen, but also for orienting effective public health strategies to mitigate disease spreading and hence for future pandemic preparedness. Laplacian-P-splines (LPS) is a fast and flexible Bayesian inference methodology anchored around Laplace approximations and P-splines. It has recently been extended to infectious disease models in the EpiLPS ecosystem (https://epilps.com) for estimation of various epidemic metrics such as the time-varying reproduction number [1], the incubation period [2] and the nowcasted incidence [3]. We highlight the role of P-splines for modeling smooth epidemic model components and illustrate the capability of EpiLPS to carry out inference through a completely sampling-free scheme that involves a negligible computational cost as compared to classic Markov chain Monte Carlo techniques. Moreover, we emphasize how the associated R package [4] can be used for analysis of real epidemic data. Finally, we discuss the benefits of EpiLPS over alternative methods and conclude with possible research extensions. [1] Gressani, O., Wallinga, J., Althaus, C., Hens, N. and Faes, C. (2022). EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. PLOS Computational Biology, 18(10): e1010618. [2] Gressani, O., Torneri, A., Hens, N. and Faes, C. (2024). Flexible Bayesian estimation of incubation times. American Journal of Epidemiology (In Press). [3] Sumalinab, B., Gressani, O., Hens, N. and Faes, C. (2023). Bayesian nowcasting with Laplacian-P-splines. MedRxiv preprint. [4] Gressani, O. (2021). EpiLPS: A fast and flexible Bayesian tool for estimating epidemiological parameters. CRAN. https://cran.r-project.org/package=EpiLPS

Statistics for non-stationary processes

Organizer: Patrice Bertail

Chair: Patrice Bertail

Room: Sala Polivalente 1.3

16:00 **Models for Science Data with Hidden Periodic Structure**
Antonio Napolitano

Abstract: Many physical phenomena are originated by the interaction of periodic phenomena with random ones. The results are processes that are not periodic but whose statistical functions are periodic functions of time. Such kind of processes are ubiquitous in science data and their hidden periodic structure can be recovered by estimating their statistical functions. In communications, radar, sonar, and telemetry, periodicities in the statistical functions are due to the modulation by random data of carriers or pulse trains. In the vibro- acoustic signals of mechanical machinery, periodicities are due to rotations of gears, belts, and bearings. In radio astronomy data, periodicities are due to the revolution and rotation of planets and pulsation of stars. In human biological signals, periodicities in statistical functions are due to heart pulsation or alternation of day and night. Hidden periodicities are present in genome sequences, diffusion processes of molecular dynamics, and signals encountered in neuroscience. If the periodic phenomenon has a regular pace, the cyclostationary and almost-cyclostationary (ACS) models can be suitably exploited for data modeling and analysis. However, in most cases, the pace is not regular and irregular cyclicity is observed in the statistical functions. In such cases, generalizations of the almost-cyclostationary model have been found to be more appropriate. These models include the classes of the generalized almost-cyclostationary (GACS) signals, of the spectrally correlated (SC) signals, and of the oscillatory almost-cyclostationary (OACS) signals. In the present talk, ACS, GACS, SC, and OACS signal models are briefly reviewed and compared. For these classes of signals, the statistical characterization and the problem of statistical function estimation are addressed. Moreover, examples of applications are illustrated.

16:30 **Harris recurrent Markov chains and nonlinear monotone cointegrated models**
Carlos Fernández, Patrice Bertail, Cecile Durot

Abstract: In this talk, we study a nonlinear cointegration-type model where the link function is a monotone function and the input process is a Harris recurrent Markov chain. Using a localization argument, we develop a non-parametric Least Square Estimator to locally estimate the link function, and under mild conditions, we show its strong consistency and obtain its rate of convergence. New results (of the Glivenko-Cantelli type) for localized null recurrent Markov chains are also presented.

17:00 **Optimal choice of bootstrap block length for periodically correlated time series**
Anna Dudek, Patrice Bertail

Abstract: We discuss the problem of choosing the optimal block length for block bootstrap methods designed for periodically correlated processes, such as the Generalized Seasonal Block Bootstrap, the Extension of Moving Block Bootstrap, and the Generalized Seasonal Tapered Block Bootstrap. We consider two estimation problems: the overall mean and the seasonal means. The obtained optimal block lengths are of the same order as for the corresponding approaches in the stationary case. They are obtained directly by minimizing the mean squared error of the corresponding bootstrap variance estimator or by exploiting the relationship between bootstrap and jackknife variance estimators. Finally, we present the results of the performed simulation study.

17:30 **Locally Stationary Spatial Processes**
Soumendra Lahiri

Abstract: The stationarity assumption is often not appropriate for modeling spatial data over large spatial domains. While the observed spatial field may be amenable to modeling by stationary random fields over smaller parts of the domain, a framework is needed to allow for (smooth) variations, both of the scaling (or the variance function) and of the spatial interactions (i.e., the spatial dependence structure) across the entire spatial domain. In this paper, we provide a formulation of locally stationary random fields that allows for such variations in the spatial covariance function. We present fairly general constructions in both the frequency and spatial domains, deriving an estimator for the local covariance based on irregularly spaced samples from the spatial process. We also establish consistency of the local covariance estimator and obtain an expansion for its mean squared error (MSE). Results from a moderately large simulation study illustrate finite sample properties of the proposed estimator.

Nonparametric methods for complex data
Organizer: Byeong Park
Chair: Byeong Park
Room: Sala Polivalente 1.5

16:00 **Inference for Changing Periodicity, Smooth Trend and Covariate Effects in Time Series**
Ming-Yen Cheng, David Siegmund, Shouxia Wang, *Lucy Xia*

Abstract: Traditional analysis of a periodic time series assumes its pattern remains the same. However, some recent empirical studies in climatology and other fields find that the amplitude may change over time, which has important implications. We develop a formal procedure to detect and estimate change-points in the periodic pattern. Often, there is also a smooth trend, and sometimes the period is unknown, with potential other covariate effects. Based on a new model that takes all of these factors into account, we propose a three-step estimation procedure to estimate them all accurately. First, we adopt penalized segmented least squares estimation for the unknown period, with the trend and covariate effects approximated by B-splines. Then, given the period estimate, we construct a novel SupF statistic and use it in binary segmentation to estimate change-points in the periodic component. Finally, given the period and change-point estimates, we estimate the entire periodic component, trend, and covariate effects. Asymptotic results for the proposed estimators are derived, including consistency of the period and change-point estimators, and the asymptotic normality of the estimated periodic sequence, trend and covariate effects. Simulation results demonstrate the appealing performance of the new method, while empirical studies highlight its advantages.

16:30 **Accelerated age-period-cohort models**
Maria Dolores Martinez-Miranda, M. Luz Gamiz, Enno Mammen, Jens Perch Nielsen

Abstract: Age-period-cohort (APC) models are important structures used to model (for example) demographic, economic, medical, behavioral and scientific output developments over time. From a nonparametric perspective, APC models can be seen as structured nonparametric density models, where the classical approach in the literature involves histogram-type estimators. In this work we develop a generalisation of APC models allowing for time acceleration in the age direction. We call the new class of models AAPC for accelerated age-period-cohort models. This new AAPC class of models comes with simple solutions to identification of the past, permissible extrapolations for the future and statistical validation: three current research challenges even in the well known APC models. The new methodology is illustrated via the important case of understanding future fertility.

17:00 **A pseudo-metric between probability distributions based on depth-trimmed regions**
Guillaume Staerman, *Pavlo Mozharovskyi*, Pierre Colombo, Stephan Clemencon, Florence D'Alche-Buc

Abstract: The design of a metric between probability distributions is a longstanding problem motivated by numerous applications in statistics and machine learning. Focusing on continuous probability distributions on the Euclidean space, we introduce a novel pseudo-metric between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces. Data depth is a nonparametric statistical tool that measures the centrality of any point in the Euclidean space with respect to a probability distribution or a data set. It is a natural median-oriented extension of the cumulative distribution function to the multivariate case. Thus, its upper-level sets - the depth-trimmed regions - give rise to a definition of multivariate quantiles. The new pseudo-metric relies on the average of the Hausdorff distance between the depth-based quantile regions with respect to each distribution. Its good behaviour with respect to major transformation groups, as well as its ability to factor out translations, are depicted. Robustness, an appealing feature of this pseudo-metric, is studied through the finite sample breakdown point. Moreover, we propose an efficient approximation method with linear time complexity with respect to the size of the data set and its dimension. The quality of this approximation as well as the performance of the proposed approach are illustrated in numerical experiments.

17:30 **Analysis in spectral domain for spatial data under fixed domain asymptotics**
Chae Young Lim, Joonho Shin, Wei-Ying Wu

Abstract: Estimation under fixed domain asymptotics for dependence of spatial processes has been actively studied, but so far most theoretical results have been investigated under specific parametric model assumptions on spatial dependence. A spectral density is one way to characterize spatial dependence for weakly stationary spatial processes. In this work, we investigate theoretical properties of a smoothed periodogram, a nonparametric estimate of spectral density, under fixed domain asymptotics. With these results, we propose a new approach to estimate tail behaviors of a spectral density.

Advances in random networks

Organizer: Marianna Pensky
Chair: Elizaveta Levina
Room: Sala Polivalente 1.4

16:00 **Signed Diverse Multiplex Networks: Clustering and Inference**
Marianna Pensky

Abstract: The paper introduces a Signed Generalized Random Dot Product Graph (SGRDPG) model, which is a variant of the Generalized Random Dot Product Graph (GRDPG), where, in addition, edges can be positive or negative. The setting is extended to a multiplex version, where all layers have the same collection of nodes and follow the SGRDPG. The only common feature of the layers of the network is that they can be partitioned into groups with common subspace structures, while otherwise matrices of connection probabilities can be all different. The setting above is extremely flexible and includes a variety of existing multiplex network models as its particular cases. The paper fulfills two objectives. First, it shows that keeping signs of the edges in the process of network construction leads to a better precision of estimation and clustering and, hence, is beneficial for tackling real world problems such as, for example, analysis of brain networks. Second, by employing novel algorithms, our paper ensures strongly consistent clustering of layers and high accuracy of subspace estimation. In addition to theoretical guarantees, both of those features are demonstrated using numerical simulations and a real data example.

16:30 **Random line graphs and edge-attributed network inference**
Avanti Athreya, Zachary Lubberts, Youngser Park, Carey Priebe

Abstract: We extend the latent position random graph model to the line graph of a random graph, which is formed by creating a vertex for each edge in the original random graph, and connecting each pair of edges incident to a common vertex in the original graph. We prove concentration inequalities for the spectrum of a line graph, as well as limiting distribution results for the largest eigenvalue and the empirical spectral distribution in certain settings. We can consistently estimate edge latent positions in a random line graph, even though such graphs are of a random size, typically have high rank, and possess no spectral gap. Our results demonstrate that the line graph of a stochastic block model exhibits underlying block structure, and in simulations, we synthesize and compare line graph-based methods against other common techniques, including tensor decompositions, for cluster recovery and edge covariate inference.

17:00 **Joint Spectral Clustering in Multilayer Degree-Corrected Stochastic Blockmodels**
Zachary Lubberts, Joshua Agterberg, Jesus Arroyo

Abstract: Modern network datasets are often composed of multiple layers, either as different views, time-varying observations, or independent sample units, resulting in collections of networks over the same set of vertices but with potentially different connectivity patterns in each network. These data require models and methods that are flexible enough to capture local and global differences across the networks, while at the same time being parsimonious and tractable to yield computationally efficient and theoretically sound solutions that are capable of aggregating information across the networks. This paper considers the multilayer degree-corrected stochastic blockmodel, where a collection of networks share the same community structure, but degree-corrections and block connection probability matrices are permitted to be different. We establish the identifiability of this model and propose a spectral clustering algorithm for community detection in this setting. Our theoretical results demonstrate that the misclustering error rate of the algorithm improves exponentially with multiple network realizations, even in the presence of significant layer heterogeneity with respect to degree corrections, signal strength, and spectral properties of the block connection probability matrices.

17:30 **Intensity Profile Projection: A Framework for Continuous-Time Representation Learning for Dynamic Networks**
Alexander Modell

Abstract: We present a new representation learning framework, Intensity Profile Projection, for continuous-time dynamic network data. Given triples (i,j,t), each representing a time-stamped (t) interaction between two entities (i,j), our procedure returns a continuous-time trajectory for each node, representing its behaviour over time. The framework consists of three stages: estimating pairwise intensity functions, e.g. via kernel smoothing; learning a projection which minimises a notion of intensity reconstruction error; and constructing evolving node representations via the learned projection. The trajectories satisfy two properties, known as structural and temporal coherence, which we see as fundamental for reliable inference. Moreover, we develop estimation theory providing tight control on the error of any estimated trajectory, indicating that the representations could even be used in quite noise-sensitive follow-on analyses. The theory also elucidates the role of smoothing as a bias variance trade-off, and shows how we can reduce the level of smoothing as the signal-to-noise ratio increases on account of the algorithm `borrowing strength' across the network.

18:00 - 19:00 Welcome Reception

Wednesday, 26 Jun

9:00 - 10:00 Contributed 2

Count data
Chair: Daniel Nevo
Room: Sala Polivalente 1.1

9:00 **The minimax risk in nonparametric testing of discrete distributions for uniformity under missing ball alternatives**
Alon Kipnis

Abstract: We study the problem of testing the goodness of fit of the data to the uniform distribution over many categories under a minimax setting in which the class of alternatives is obtained by the removal of an l_p ball of a radius r around the uniform rate sequence. We provide an expression describing the sharp asymptotic of the minimax risk in terms of these parameters as N goes to infinity. Our result settles an open question related to an extensive line of work over the last two decades. Furthermore, it allows the comparison of the many estimators previously proposed for this problem at the constant level, rather than at the rate of convergence of the risk or the scaling order of the sample complexity. The minimax test mostly relies on collisions in the very small sample limit but behaves like the chisquared test for moderate and large sample sizes. Empirical studies over a range of problem parameters show that the asymptotic estimate of the minimax risk is accurate in finite samples and that the asymptotic minimax test is significantly better than the chisquared test or a test that only uses collisions. The analysis relies on the adaptation of relatively known methods in nonparametric normal signal detection to the Poisson setting. Specifically, the equivalence between the minimax setting and a Bayesian setting, and the characterization of the least favorable prior by reducing a multi-dimensional optimization problem to a one-dimensional problem.

9:20 **Semiparametric test for overdispersed count data**
Stefano Bonnini, Michela Borghesi

Abstract: When the response of a regression model is a count variable, the classic methodological solution is based on the Generalized Linear Model approach. To test the model adequacy, under some conditions, for example in the presence of overdispersed data, the typical solutions based on the Likelihood Ratio Test, in the framework of the Poisson regression and of the Negative Binomial regression, are not performant and not suitable. In such conditions, a semiparametric approach based on the combination of permutation tests on the significance of the coefficient estimates, by using the maximum likelihood estimates as test statistics of such permutation tests, is proposed. The good performance in terms of power behavior and the practical usefulness of the proposal are proved through a comparative Monte Carlo simulation study and the application of the method to an interesting case study.

9:40 **Multiple change-point detection in a Poisson process**
Emilie Lebarbier

Abstract: Change-point detection aims at discovering behavior changes lying behind time sequences data. In this talk, we investigate the case where the data come from an inhomogenous Poisson process. We present an offline multiple change-point detection methodology based on minimum contrast estimator. The main difficulty concerns the estimation of the position of the change-points. Indeed, the classical contrasts (minus log-likelihood or least-square) are not convex and nor even continous with respect to these parameters. This problem has been already observed and solved in the literature for the detection of one change-point . We show that their idea can be extended to the multiple change-point detection for a general class of contrasts. The optimization problem is thus reduced to a discrete optimization problem which can be solved using a well-efficient algorithm known and used for the detection problem with discret observations cases. Besides, we select the appropriate number of change-points through a cross-validation procedure which is really convenient here due to the nature of the Poisson process. Through experiments on simulated and realworld datasets, we show the interest of the proposed method, which is implemented in the CptPointProcess R package.

Extremes
Chair: Armelle Guillou
Room: Sala Polivalente 1.3

9:00 **Extremal behaviour and convergence rates for sample-based geometric quantiles and half space depths**
Marie Kratz

Abstract: We consider the empirical versions of geometric quantile and halfspace depth, and study their extremal behaviour as a function of the sample size. The objective of this study is to establish connection between the rates of convergence and tail behaviour of the corresponding underlying distributions. The intricate interplay between the sample size and the parameter driving the extremal behaviour forms the main result of this analysis. In the process, we also fill certain gaps in the understanding of population versions of geometric quantile and halfspace depth. This is a joint work with S. Singha and S. Vadlamani.

9:20 **Spatio-temporal model for the occurrence of extreme events and inference on their extent**
Ana C. Cebrian

Abstract: Modeling extreme events (EE) typically involves identifying exceedances of a suitable threshold within the response series. In this study, we present a spatio-temporal model capable of predicting both the occurrence and characteristics, such as intensity and duration, of EEs at any location within the study region. Our model utilizes a two-state framework for EEs, incorporating local thresholds to model the response series. This approach is specifically employed to model extreme heat events (EHE) in daily temperature series. The model switches between two observed states: one defining extreme heat days (above the temperature threshold) and the other defining non-extreme heat days. This two-state structure accommodates temporal dependence while allowing parameters controlling spatial dependence to vary between the two states. Transition probabilities are determined by a two-state Markovian switching model, with each sub-model incorporating seasonal terms, covariates, and intercepts modeled as Gaussian processes. Additionally, we introduce a formal definition of the spatial extent of an extreme event, illustrating how it can be calculated using the output from the previous model. For a specified region and day, the spatial extent is computed as the block average of indicator functions over the region. Our risk assessment focuses on Aragón (NE Spain), and we conduct decade-wise comparisons to reveal evidence of an increasing extent over time, serving as an indicator of the effects of global warming.

9:40 **Risk Assessment using a Semi-Parametric Approach**
Ayana Mateus, Frederico Caeiro

Abstract: In the domain of Statistics of Extremes, accurately estimating the extreme value index is pivotal for tail inference. This study focuses on Pareto-type models and It employs a semi-Parametric framework. We introduce a novel class of estimators grounded on weighted moments, offering robust estimation of the extreme value index. These estimators provide valuable insights into tail behavior and facilitate the estimation of other tail parameters, including return periods and extreme quantiles. Those parameters play a critical role in assessing the frequency and severity of rare events, crucial for risk assessment, infrastructure planning, and disaster preparedness.

Functional data analysis 2
Chair: Annika Betken
Room: Grande Auditorio

9:00 **Measuring dependence between a scalar response and a functional covariate**
Daniel Strenger, Siegfried Hörmann

Abstract: We explore a dependence coefficient between scalar responses and functional covariates. It turns out that the limiting behavior of the sample version of this coefficient is quite delicate, as it crucially depends on the nearest neighbor structure of the covariate sample. Essentially, one needs an upper bound for the maximal number of points which share the same nearest neighbor. While a deterministic bound exists for multivariate data, this is no longer the case in infinite dimensional spaces. To our surprise, relatively little seems to be known about properties of the nearest neighbor graphs in high-dimensional or functional random sample, and hence the main contribution of this paper is to advise a way how to overcome this problem. An important application of our theoretical results is a test for independence between scalar responses and functional covariates. This test is consistent against all fixed alternatives.

9:20 **Directional regularity: Achieving faster rates of convergence in multivariate functional data**
Sunny Wang, *Omar Kassi*

Abstract: We consider a new notion of regularity, called directional regularity, which is relevant for a wide range of applications involving multivariate functional data. We show that among the class of seemingly isotropic processes, a subset of anisotropic processes exist. Faster rates of convergence may thus be obtained by adapting to their directional regularity through a change of basis. Algorithms are constructed for the estimation and identification of the directional regularity, made possible due to the unique replication nature of functional data, with accompanying non-asymptotic theoretical guarantees provided. A novel simulation algorithm for anisotropic processes is designed to evaluate the numerical accuracy of our directional regularity algorithm. Simulation results demonstrate the good finite sample properties of our estimator. Applications which elucidate the concrete benefits of our methodology are discussed and illustrated.

9:40 **Functional relevance based on continuous Shapley value**
Pedro Delicado, Cristian Pachón-García

Abstract: The presence of machine learning models in many facets of our lives has multiplied in recent years. Often, improvements in predictive efficiency come at the cost of increasing their complexity, which is why they are referred to as "black boxes". The opacity of certain algorithms has led to a growing demand to understand how and why they make their decisions. In response, a whole literature has recently emerged ("Interpretable Machine Learning" or "eXplainable Artificial Intelligence") whose goal is to make automatic algorithms transparent and interpretable. Among the interpretability methods, special attention has been given to those that are model agnostic (can be applied to any predictive model) and global (they measure the relevance/importance of each variable over the entire dataset). Multicolinearity among predictors makes it difficult to assign individual relevance measures to them. To overcome this problem, Lipovesky and Conklin (2001) proposed to adapt Shapley value imputation (a concept from cooperative games theory) to measure regressors' importance. Consider a scalar-on-function regression problem, where the goal is to predict a scalar response from a functional predictor. Several predictive models have been proposed in the Functional Data Analysis literature: linear and nonlinear models, parametric and nonparametric, among others. In addition, other proposals have come from the machine learning literature: Support Vector Machine regression can be adapted to functional data, and versions of Neural Networks for functional data have recently appeared (Heinrichs et al., 2023). The above scalar-on-function predictive methods are generally difficult to interpret because it is hard to identify which features of the functional predictors are more important in computing the predicted values. In this work, we extend relevance measures based on the Shapley value from multivariate to functional predictors by adapting concepts from the continuous games literature. Our proposals are illustrated by a simulation study and several real data applications.

Smoothing methods
Chair: José E. Chacón
Room: Sala Polivalente 1.2

9:00 **Bayesian wavelet regression using nonlocal prior mixtures and novel parameterizations**
Nilotpal Sanyal

Abstract: We propose a Bayesian nonparametric regression method using wavelets. For the wavelet coefficients, we consider a mixture prior based on nonlocal prior mixtures. This is motivated by the observation that different nonlocal priors provide better support for smaller and larger effect sizes in biological data such as the genome-wide association study data. Specifically, we consider a prior that is a mixture of a point mass, a method of moments (MOM) prior and an inverse method of moments (IMOM) prior. We specify the mixture probabilities of the nonlocal priors through novel parametric forms that resemble the distribution functions of generalized logistic, hyperbolic secant, and generalized normal distributions, and uses several hyperparameters. In addition, for the variance components of the nonlocal priors, we consider a novel double-exponential decay. Following an empirical Bayes approach, we estimate the hyperparameters from the data and develop posterior inference conditioning on the hyperparameter estimates making use of the Laplace's method. We perform extensive simulation analyses using simulated datasets that combine typical spatial variability features in various proportions to assess and compare the performances of our proposed method under various parameterizations, and provide guidelines for specific usage. Further, real data applications demonstrate the utility and flexibility of our proposed method.

9:20 **The Effective Degrees of Freedom in Kernel Density Estimation**
Alex Trindade, Sofia Guglielmini, Igor Volobouev

Abstract: The quest for a formula that satisfactorily measures the effective degrees of freedom in kernel density estimation (KDE) is a long standing problem with few solutions. Starting from the oracle orthogonal polynomial sequence (OPS) expansion for the underlying density, we show how convolution with the kernel leads to a new OPS with respect to which one may express the resulting KDE. The expansion coefficients of the two OPS systems can then be related via a kernel sensitivity, or smoothing, matrix, and this then naturally leads to a definition of effective parameters by taking the trace of a symmetrized version. The resulting effective degrees of freedom (EDoF) formula is akin to that proposed by Loader (1999) for local likelihood and more recently by McCloud & Parmeter (2020) for KDE. Meaningfully, four different formulations are shown to result in the same EDoF. Empirical estimates are proposed, and their asymptotic properties worked out through influence functions. The methodology offers the tantalizing possibility of applying covariance penalty based model selection procedures like AIC & BIC in the heretofore forbidden realm of KDE.

9:40 **Statistical modelling of the firing activity of grid cells using local polynomial kernel smoothing methods**

Rida Ayyaz, Ioannis Papastathopoulos, Mathew Nolan

Abstract: In this study, we investigate non-parametric kernel smoothing methods for estimating the intensity function of a point pattern. In particular, we show how to smooth the likelihood function of a non-homogeneous Poisson process by assigning weights to all contributions in the likelihood, including the ones that come from the observed events as well as those coming from the part of the domain where no events are observed. Subsequently, we use this weighted likelihood to introduce local polynomial regression models for the logarithm of the intensity function of a point pattern and show how this framework leads to traditional estimators alongside novel local log-linear estimators of the intensity function. Additionally, we implement cross-validation methods to estimate the bandwidth and discuss why commonly used methods such as leave-one-point-out cross-validation only assess the local scale variations in predictions and thus, may not provide accurate choices for the bandwidth. This motivates us to investigate, under this new setting, an alternative leave-group-out cross-validation approach for learning the bandwidth parameter in a data-driven manner. We use bootstrap to quantify uncertainty in estimation and showcase our methods via an application to spike-train data from grid cells, a class of neurons located in the medial entorhinal cortex of the brain, playing a key role in encoding spatial information and how animals learn to navigate in their environment.

Survival Analysis 2

Chair: Rebecca Betensky

Room: Pequeno Auditorio

9:00 **Increasing odds ratio, testing and applications**

Paulo Eduardo Oliveira, Idir Arab, Tommaso Lando

Abstract: Distributions with increasing odds ratio define an interesting and wide class, that may be used for benchmarking purposes for adverse aging properties. We discuss some characterizations of this class, proposing applications to the construction of tolerance bounds. Moreover, we prove that a test based on the greatest convex minorant has good properties, providing a simulation based description of its distribution.

9:20 **Quantile modelling under dependent censoring**

Myrthe D'Haen, Ingrid Van Keilegom, *Anneleen Verhasselt*

Abstract: In survival analysis under random right censoring, one may observe a censoring time C for some values rather than the survival time T. Often, such censoring is dealt with under an independence assumption on T and C, given the covariate X. However, in some cases, this may not be a very realistic assumption; by taking care of the possible dependency, inference on the survival time could be handled more accurately. In this research, this inference for T is done with a focus on quantile regression, but some broader regression results are obtained as a by-product. In order to capture any dependence, the quantile model for T is derived from a bivariate copula model for (T, C). For this copula model, a flexible copula parameter is taken to deal with the practice of often unknown association. It comes at the cost of marginals that are necessarily fully parametric, but this can be overcome by considering the family of so-called enriched asymmetric Laplace (EAL) distributions for T. While preserving the parametric character, they enable introducing sufficient modelling flexibility by means of Laguerre orthogonal polynomials.

9:40 **Testing for sufficient follow-up in survival data with a cure fraction**

Tsz Pang Yuen, Eni Musta

Abstract: In order to estimate the proportion of 'cured' subjects who will never experience failure, a sufficiently long follow-up period is required. Several statistical tests have been proposed in the literature for testing the assumption of sufficient follow-up, such as the tests in Maller and Zhou (1992, 1994) and in Maller et al. (2023). These tests consider the follow-up as sufficient when the study duration is longer than the support of the survival times for the uncured subjects. However, for practical purposes, the follow-up would be considered sufficiently long if the probability for the event to happen after the end of the study is very small. Based on this observation, we formulate a more relaxed notion of 'practically' sufficient follow-up characterized by the quantiles of the distribution and develop a novel nonparametric statistical test. The proposed method relies mainly on the assumption of a non-increasing density function in the tail of the distribution. The test is then based on a shape constrained density estimator such as the Grenander or the kernel smoothed Grenander estimator and a bootstrap procedure is used for computation of the critical values. The performance of the test is investigated through a simulation study, and the method is illustrated on a breast cancer data.

Time series 2

Chair: Sophie Dabo

Room: Sala Polivalente 1.4

9:00 **Bootstrap inference for group factor models**
Benoit Perron, Silvia Gonçalves, Julia Koh

Abstract: Recently, Andreou, Gagliardini, Ghysels, and Rubin (2019) have proposed a test for common factors based on canonical correlations between factors estimated separately from each group. We propose and theoretically justify a simple bootstrap test that avoids the need to estimate the bias and variance of the canonical correlations explicitly. We verify these conditions for a wild bootstrap scheme similar to the one proposed in Gonçalves and Perron (2014). We also propose a more general bootstrap scheme that is valid under general cross-sectional and serial dependence. Simulation experiments show that this bootstrap approach leads to rejection rates closer to the nominal level in all of our designs compared to the asymptotic framework.

9:20 **Distribution-free prediction intervals from interval score optimized pairs of nonparametric regression quantiles**
Harry Haupt, *Joachim Schnurbus*, Ida Bauer

Abstract: We propose a distribution-free method for probabilistic prediction of trending seasonal time series over time horizons of up to one seasonal cycle. Prediction intervals are computed by a pair of nonparametric regression quantiles using a local-linear approach. The bandwidth selection is based on minimizing the average of the training-sample interval score, balancing the trade-off between coverage and length. We propose a simple selection rule to modify the prequential training algorithm to adjust this trade-off for better coverage of the estimated test-sample prediction intervals. We study the proposed interval prediction method for short-term predictions of the monthly and quarterly U.S. industrial production index over the last 50 years. For a comprehensive picture of the interval prediction performance, we extend this study to several hundred hourly, daily, monthly, and quarterly time series from the M4 competition.

9:40 **Pointwise spectral density estimation under local differential privacy**
Karolina Klockmann, Tatyana Krivobokova, Cristina Butucea

Abstract: We propose a new interactive locally differentially private mechanism for estimating Hölder-smooth spectral density functions of stationary Gaussian processes. Anonymization is achieved through two-stage truncation and subsequent Laplace perturbation. In particular, we show that our method achieves a pointwise L2-rate with a dependency of only a^2 on the privacy parameter a. This rate stands in contrast to the results of (Kroll, 2024), who proposed a non-interactive mechanism for spectral density estimation and showed a dependency of a^4 on the privacy parameter for the uniform L2-rate.

10:00 - 11:00 Keynote Talk 1
Chair: Igor Pruenster
Room: Grande Auditorio

10:00 **Causality-inspired Statistical Machine Learning**
Peter Bühlmann

Abstract: Reliable, robust and interpretable machine learning is a big emerging theme in data science and statistics, complementing the development of pure black box prediction algorithms. New connections between distributional robustness, external validity and causality provide methodological paths for improving the reliability and understanding of machine learning algorithms, with wide-ranging prospects for various applications.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk 1
Chair: Sophie Dabo
Room: Grande Auditorio

11:30 **Provably Fast and Accurate Estimation of Neural Nets**
Andrew R. Barron

Abstract: Recent developments in fast algorithms for relaxed greedy training of artificial neural networks are presented. Both guaranteed optimization procedures and greedy Bayes sampling procedures with rapid mixing are explored. Along with these algorithmic developments we present corresponding bounds for statistical risk and for online learning regret for predictions based on these neural network fits. This is based on joint work with Curtis McDonald of Yale University and Jason Klusowski of Princeton University.

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 3

Nonstationary processes: theory and applications

Organizer: Anna Dudek
Chair: Anna Dudek
Room: Sala Polivalente 1.3

13:30 **Bayesian nonparametric spectral analysis of locally stationary processes**
Yifu Tang, *Claudia Kirch*, Jeong Eun Lee, Renate Meyer

Abstract: Many real-world phenomena can well be modelled by locally stationary time series with a slowly changing dependency structure. For such time series the estimation of the time-varying spectral density is of particular interest. In this talk we propose a Bayesian nonparametric method for this purpose based on a suitable dynamic Whittle likelihood for locally stationary time series in combination with a Bernstein-Dirichlet process prior. We prove sup-norm posterior consistency and obtain L2-norm posterior contraction rates. Additionally, this methodology enables model selection between stationarity and non-stationarity based on the Bayes factor. The good finite-sample performance of the method is indicated by means of a simulation study and applications to real-life data-sets.

14:00 **Spectral analysis and subsampling for spectrally correlated processes**
Anna E. Dudek, *Bartosz Majewski*

Abstract: Spectrally correlated processes are harmonizable processes characterized by spectral measures concentrated within a countable union of curves. Our study focuses on the spectral analysis of spectrally correlated processes, particularly in scenarios where these support curves take the form of lines with possible non-unit slopes. This class of spectrally correlated processes finds practical application in, for example, the problem of locating moving sources like aircraft, rockets, or other hostile jamming emitters emitting communication signals. We propose a frequency-smoothed periodogram along the support line as the spectral density function estimator. We prove its mean-square consistency and derive its asymptotic distribution. Based on these results, we discuss the asymptotic properties of the introduced coherence estimator. Additionally, we formulate a subsampling technique tailored for spectral analysis of spectrally correlated processes.

14:30 **Trend estimation in a class of explosive count time series**
Anne Leucht, Michael Neumann

Abstract: We consider a log-linear model for non-stationary count time series with an exploding trend. The model is designed such that the conditional mean and the conditional standard deviation are of the same order of magnitude which is comparable to classical GARCH models. Note that this type of (conditional) overdispersion is not captured by the usual Poisson INGARCH models. We fit a polynomial trend using an ordinary least squares approach and show asymptotic normality of our estimator at a non-standard rate. Since the limiting variance depends on unknown parameters, we propose a variant of the dependent wild bootstrap to derive confidence intervals. We provide a result on its asymptotic validity and illustrate the finite sample performance of the proposed methods by a simulation study. This talk is based on joint work with Michael H. Neumann (Friedrich-Schiller-Universität Jena).

15:00 **Nonparametric hypothesis testing for the structure of spectrum of nonstationary processes**
Jean-Marc Freyermuth

In this talk, we present a method developed in Aston et al. (2019a) for studying structural characteristics of multidimensional functions such as the spectrum of spatio-temporal processes. We consider the multidimensional Gaussian white noise model with an anisotropic estimand. Using the relation between the Sobol decomposition and the geometry of the hyperbolic wavelet basis, we can build test statistics for any of the Sobol functional components. We assess the asymptotic minimax performance of these test statistics and show that they are optimal in presence of anisotropy with respect to the newly determined minimax separation rates. An appropriate combination of these test statistics allows testing some general structural characteristics. Numerical experiments show the potential of our method for studying the spectrum of spatio-temporal processes. As a conclusion, we briefly discuss extension of this work for testing the structure of spectral characteristics of harmonizable processes (Dehay et al. (2024)) and toward a novel theoretical approach to assess the performance of signal detection procedures (Autin et al (2019b)). [1] Aston, J., Autin, F., Claeskens, G., Freyermuth, J-M. (2019a). Minimax optimal procedures for testing the structure of multidimensional functions. Applied Computational Harmonic Analysis, 46(2), 288-311. [2] Dehay, D., Dudek, A.E., Freyermuth, J-M. Spectral characteristics of Harmonizable VARMA processes. 2023. Preprint. [3] Autin, F., Clausel, M., Freyermuth, J-M., Marteau, C. (2019b). Maxiset point of view for signal detection in inverse problems. Mathematical Methods of Statistics, 28, 228-242.

Theory and methods in Bayesian nonparametrics: recent advances

Organizer: Antonio Lijoi
Chair: Igor Pruenster
Room: Sala Polivalente 1.4

13:30 **Functional connectivity across the human subcortical auditory system using an autoregressive matrix-Gaussian copula graphical model approach with partial correlations**
Noirrit Kiran Chandra, Kevin Sitek, Bharath Chandrasekaran, Abhra Sarkar

Abstract: The auditory system comprises multiple subcortical brain structures that process and refine incoming acoustic signals along the primary auditory pathway. Due to technical limitations of imaging small structures deep inside the brain, most of our knowledge of the subcortical auditory system is based on research in animal models using invasive methodologies. Advances in ultra-high field functional magnetic resonance imaging (fMRI) acquisition have enabled novel non-invasive investigations of the human auditory subcortex, including fundamental features of auditory representation such as tonotopy and periodotopy. However, functional connectivity across subcortical networks is still underexplored in humans, with ongoing development of related methods. Traditionally, functional connectivity is estimated from fMRI data with full correlation matrices. However, partial correlations reveal the relationship between two regions after removing the effects of all other regions, reflecting more direct connectivity. While most existing methods for learning conditional dependency structures based on partial correlations assume independently and identically Gaussian distributed data, fMRI data exhibit significant deviations from Gaussianity as well as high temporal autocorrelation. In this paper, we developed an autoregressive matrix-Gaussian copula graphical model approach to estimate the partial correlations and thereby infer the functional connectivity patterns within the auditory system while appropriately accounting for autocorrelations between successive fMRI scans. Our results are highly stable when splitting the data in halves according to the acquisition schemes and computing partial correlations separately for each half of the data, as well as across cross-validation folds. In contrast, full correlation-based analysis identified a rich network of interconnectivity that was not specific to adjacent nodes along the pathway. Overall, our results demonstrate that unique functional connectivity patterns along the auditory pathway are recoverable using novel connectivity approaches and that our connectivity methods are reliable across multiple acquisitions.

14:00 **Constrained Dirichlet Processes and Moment Condition Models**
Jaeyong Lee

Abstract: In this talk, we consider the constrained Dirichlet process (cDP), the conditional distribution of the Dirichlet process when a functional of a random distribution is given. We specifically apply the cDP to the moment condition model. This model is a nonparametric model in which the finite dimensional parameter of interest is defined as a solution to a functional equation of the distribution. We derive both the posterior distribution of the parameter of interest and that of the underlying distribution itself. We investigate the properties of the moment condition model with cDP and propose an algorithm for the posterior inference.

14:30 **Distances on random probability measures**
Marta Catalano

Abstract: Random probabilities are a key component to many nonparametric methods in Statistics and Machine Learning. To quantify comparisons between different laws of random probabilities several works are starting to use the elegant Wasserstein over Wasserstein distance. In this talk we show that the infinite-dimensionality of the space of probabilities drastically deteriorates its sample complexity, which is slower than any polynomial rate in the sample size. We thus propose a new distance that preserves many desirable properties of the former while achieving a parametric rate of convergence. In particular, our distance 1) metrizes weak convergence; 2) can be estimated numerically through samples with low complexity; 3) can be bounded analytically from above and below. The main ingredient are integral probability metrics, which lead to the name hierarchical IPM. This is joint work with Hugo Lavenant.

15:00 **Bayesian Nonparametrics with the Martingale Posterior**
Edwin Fong

Abstract: While the prior distribution is the usual starting point for Bayesian uncertainty, recent work has reframed Bayesian inference as the predictive imputation of missing observations. In particular, the martingale posterior distribution arises when the Bayesian model is a chosen sequence of predictive distributions on future observables, which then induces a posterior distribution on the parameter of interest without the need for a likelihood and prior. This generalization broadens the range of models one can use for nonparametric Bayesian inference, and offers advantages in computation and flexibility. In this talk, we introduce the framework and present some recent advances, with a focus on predictive Bayesian nonparametric methodologies.

Estimation and testing problems with survival data

Organizer: Jacobo de Uña-Álvarez
Chair: Jacobo de Uña-Álvarez
Room: Sala Polivalente 1.5

13:30 **Surviving the multiple testing problem: RMST-based tests in general factorial designs**
Merle Munko, Marc Ditzhaus, Dennis Dobler, Jon Genuneit

Abstract: Several methods in survival analysis are based on the proportional hazards assumption. However, this assumption is very restrictive and often not justifiable in practice. Therefore, effect estimands that do not rely on the proportional hazards assumption, such as the restricted mean survival time (RMST), are highly desirable in practical applications. The RMST is defined as the area under the survival curve up to a prespecified time point and, thus, summarizes the survival curve into a meaningful estimand. For two-sample comparisons based on the RMST, there is an inflation of the type-I error of the asymptotic test for small samples and, therefore, a two-sample permutation test has already been developed. The first goal is to further extend the permutation test for general factorial designs and general contrast hypotheses by considering a Wald-type test statistic and its asymptotic behavior. Additionally, a groupwise bootstrap approach is considered. In a second step, multiple tests for the RMST are developed to infer several null hypotheses simultaneously. Hereby, the asymptotically exact dependence structure between the local test statistics is incorporated to gain more power. The small sample performance of the proposed global and multiple testing procedures is analyzed in simulations and finally illustrated by analyzing a real data example.

14:00 **Bivariate dependent censoring with covariates**
Noël Veraverbeke

Abstract: We consider a bivariate vector (T1, T2) of lifetimes subject to right censoring by a vector (C1, C2). There is also a random covariate X and the purpose is to estimate the conditional joint survival function of (T1, T2), given that the covariate X takes some value x. A common assumption is that the vectors (T1, T2) and (C1, C2) are conditionally independent, given X. But this assumption is not always satisfied in practical situations. A solution that has been proposed in the univariate case is to specify a known copula for the dependence between lifetime and censoring time. We explore how far we get with this idea in the bivariate case. In the second part we study an important particular case: the one-component censoring model. In this model, the component T1 is always fully observed and the second component T2 is subject to right censoring by C2. W e assume that T2 and C2 are linked by a known conditional copula. This, together with a 'Stute' type assumption on the conditional probability of censoring leads to a solution of our estimation problem.

14:30 **A fully parametric model for non-proportional hazards survival analysis**
María del Carmen Pardo, María del Mar Fenoy, Narayanaswamy Balakrishnan

Abstract: Despite the Cox's proportional hazards (PH) model is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables, it is being realised increasingly that not all survival data obey the PH assumption. A generalized time dependent logistic (GTDL) model was introduced by MacKenzie (1996) as an alternative to the popular Cox proportional-hazards model. Our proposal is a family which contains as particular case this model. The loglikelihood approach is used to estimate the model parameters. We perform a simulation study to evaluate the finite-sample performance of these estimators. Finally, this new family of models is illustrated with gastrointestinal cancer data. [1] MacKenzie, G. (1996). Regression models for survival data: the generalized time-logistic family.Statistician, 45, 21–34.

15:00 **Estimation and regression for sequentially-truncated data**
Rebecca Betensky, Jing Qian, Erik Parner, Morten Overgaard

Abstract: In observational cohort studies with complex sampling schemes, truncation arises when the time to event of interest is observed only when it falls below or exceeds another random time, i.e., the truncation time. In more complex settings, observation may require a particular ordering of event times; we refer to this extension of the traditional paradigm as sequential truncation and partial sequential truncation. I first describe nonparametric and semiparametric maximum likelihood estimators for the distribution of the event time of interest in the setting of these truncation settings. I then describe methods for regression modeling in this complex setting using the tool of pseudo-observations (PO). PO's are jackknife-like constructs that estimate an individual's contribution to an estimand. They are attractive in this setting because they obviate the need to directly account for the sequential truncation in the regression model of interest. Importantly, they may not be used when the truncation depends on the covariates that explain the time-to-event of interest; in this case a modified PO approach is available. We consider both the Cox and accelerated failure time (AFT) models. We evaluate our approach in simulation studies and in application to an Alzheimer's cohort study.

Large scale semi-parametric inference
Organizer: Omiros Papaspiliopoulos
Chair: Omiros Papaspiliopoulos
Room: Sala Polivalente 1.2

13:30 **Partially factorized variational inference for high-dimensional mixed models**
Max Goplerud, Omiros Papaspiliopoulos, Giacomo Zanella

Abstract: While generalized linear mixed models (GLMMs) are a fundamental tool in applied statistics, many specifications - such as those involving categorical factors with many levels or interaction terms - can be computationally challenging to estimate due to the need to compute or approximate high-dimensional integrals. Variational inference (VI) methods are a popular way to perform such computations, especially in the Bayesian context. However, naive VI methods can provide unreliable uncertainty quantification. We show that this is indeed the case in the GLMM context, proving that standard VI (i.e. mean-field) dramatically underestimates posterior uncertainty in high-dimensions. We then show how appropriately relaxing the mean-field assumption leads to VI methods whose uncertaintyquantification does not deteriorate in high-dimensions, and whose total computational cost scales linearly with the number of parameters and observations. Our theoretical and numerical results focus on GLMMs with Gaussian or binomial likelihoods, and rely on connections to random graph theory to obtain sharp high-dimensional asymptotic analysis. We also provide generic results, which are of independent interest, relating the accuracy of variational inference to the convergence rate of the corresponding coordinate ascent variational inference (CAVI) algorithm for Gaussian targets. Our proposed partially-factorized VI (PF-VI) methodology for GLMMs is implemented in the R package vglmer. Numerical results with simulated and real data examples illustrate the favourable computation cost versus accuracy trade-off of PF-VI.

14:00 **Penalized likelihood estimation and inference in high-dimensional logistic regression**
Ioannis Kosmidis, Philipp Sterzinger

Abstract: In recent years, there has been a surge of interest in estimators and inferential procedures that exhibit optimal asymptotic properties in high-dimensional logistic regression when the number of covariates grows proportionally as a fraction kappa in (0,1) of the number of observations. In this seminar, we focus on the behaviour of a class of maximum penalized likelihood estimators, employing the Diaconis-Ylvisaker prior as the penalty. Building on advancements in approximate message passing, we analyze the aggregate asymptotic behaviour of these estimators when covariates are normal random variables with arbitrary covariance. This analysis enables us to eliminate the estimators' persistent asymptotic bias through straightforward rescaling for any value of the prior hypertuning parameter. Moreover, we derive asymptotic pivots for constructing inferences, including adjusted Z-statistics and penalized likelihood ratio statistics. Unlike the maximum likelihood estimate, which only asymptotically exists in a limited region on the plane of kappa versus signal strength, the maximum penalized likelihood estimate always exists and is directly computable via maximum likelihood routines. As a result, our asymptotic results remain valid even in regions where existing maximum likelihood results are not obtainable, with no overhead in implementation or computation. The estimators' dependency on the prior hyper-parameter facilitates the derivation of estimators with zero asymptotic bias and minimal mean squared error. We will explore these estimators' shrinkage properties, substantiate our theoretical findings with simulations and applications, and present evidence for conjectures with non-normal covariate distributions.

14:30 **On the role of parametrization in models with a misspecified nuisance component**
Heather Battey

Abstract: I will speak about some recent work, coauthored with Nancy Reid, on inference for a parameter of interest in models that share a common interpretation for that parameter but that may differ appreciably in other respects. We study the general structure of models under which the maximum likelihood estimator of the parameter of interest is consistent under arbitrary misspecification of the nuisance part of the model. A specialization of the general results to matched-comparison and two-groups problems gives a more explicit and easily checkable condition in terms of a new notion of symmetric parametrization, leading to an appreciable broadening and unification of existing results in those problems. The role of a generalized definition of parameter orthogonality is highlighted, as well as connections to Neyman orthogonality. If time permits, the issues involved in obtaining inferential guarantees beyond consistency will be briefly discussed.

15:00 **Empirical partially Bayes multiple testing and compound chi-square decisions**
Nikolaos Ignatiadis, Bodhisattva Sen

Abstract: A common task in high-throughput biology is to screen for associations across thousands of units of interest, e.g., genes or proteins. Often, the data for each unit are modeled as Gaussian measurements with unknown mean and variance and are summarized as per-unit sample averages and sample variances. The downstream goal is multiple testing for the means. In this domain, it is routine to "moderate'' (that is, to shrink) the sample variances through parametric empirical Bayes methods before computing p-values for the means. Such an approach is asymmetric in that a prior is posited and estimated for the nuisance parameters (variances) but not the primary parameters (means). Our work initiates the formal study of this paradigm, which we term "empirical partially Bayes multiple testing.'' In this framework, if the prior for the variances were known, one could proceed by computing p-values conditional on the sample variances---a strategy called partially Bayes inference by Sir David Cox. We show that these conditional p-values satisfy an Eddington/Tweedie-type formula and are approximated at nearly-parametric rates when the prior is estimated by nonparametric maximum likelihood. The estimated p-values can be used with the Benjamini-Hochberg procedure to guarantee asymptotic control of the false discovery rate. Even in the compound setting, wherein the variances are fixed, the approach retains asymptotic type-I error guarantees.

Recent advances in time series and functional data analysis
Organizer: Alexander Aue
Chair: Jens-Peter Kreiss and Efstathios Paparoditis
Room: Sala Polivalente 1.1

13:30 **Intrinsic and Extrinsic Graphical Models for Functional Data**
Victor Panaretos

Abstract: Graphical models allow us to distinguish direct and indirect associations in data. They can be used as models or they can be the targets of inference, and have been extensively studied in many contexts, the most recent emphasis being in high-dimensional statistics. For the most part, one considers a collection of random vectors indexed by a finite or discrete set. The purpose of this talk is to explore what happens when we are interested in the associations within (intrinsic) and between (extrinsic) continuous time stochastic processes. In the first case, we are dealing with graphical models with uncountable rather than discrete index set. In the latter case, we are dealing with graphical models for finite collections of function-valued random elements. Such settings introduce conceptual challenges owing to the lack of infinite-dimensional analogues of familiar algebraic tools, such as matrix inverses and determinants. It will turn out that the two problems share commonalities but also notable differences. Based on joint work with Kartik Waghmare (EPFL) and Tomas Masak (EPFL).

14:00 **Integrative analysis of Riemannian and high-dimensional data**
Eardi Lila, James Buenfil

Abstract: In this talk, I will present a novel statistical method for the integrative analysis of Riemannian-valued functional data and high-dimensional data. The motivating application is the exploration of the dependence structure between a subject's dynamic functional connectivity -- represented by a temporally indexed collection of positive definite covariance matrices -- and high-dimensional data representing lifestyle, demographic, and psychometric measures. To this purpose, we employ a functional regression-based reformulation of canonical correlation analysis that allows us to control the complexity of the connectivity canonical directions within a Riemannian framework, using tangent space principal components analysis, and that of the high-dimensional canonical directions via a sparsity-promoting penalty. The proposed method shows improved empirical performance over alternative approaches. Its application to data from the Human Connectome Project reveals a dominant mode of covariation between dynamic functional connectivity and lifestyle, demographic, and psychometric measures. This mode aligns with results from static connectivity studies but reveals a unique temporal non-stationary pattern that such studies fail to capture.

14:30 **A statistical framework for analyzing shape in a time series of random geometric objects**
Anne van Delft, Andrew J. Blumberg

Abstract: We introduce a new framework to analyze shape descriptors that capture the geometric features of an ensemble of point clouds. At the core of our approach is the point of view that the data arises as sampled recordings from a metric space-valued stochastic process, possibly of nonstationary nature, thereby integrating geometric data analysis into the realm of functional time series analysis. We focus on the descriptors coming from topological data analysis. Our framework allows for natural incorporation of spatial-temporal dynamics, heterogeneous sampling, and the study of convergence rates. Further, we derive complete invariants for classes of metric space-valued stochastic processes in the spirit of Gromov, and relate these invariants to so-called ball volume processes. Under mild dependence conditions, a weak invariance principle in D([0,1] x [0,R]) is established for sequential empirical versions of the latter, assuming the probabilistic structure possibly changes over time. Finally, we use this result to introduce novel test statistics for topological change, which are distribution free in the limit under the hypothesis of stationarity.

15:00 **Prediction of Singular VARs and an Application to Generalized Dynamic Factor Models**
Siegfried Hörmann, Gilles Nisol

Abstract: Vector autoregressive processes (VARs) with innovations having a singular covariance matrix (in short singular VARs) appear naturally in the context of dynamic factor models. Estimating such a VAR is problematic, because the solution of the corresponding equation systems is numerically unstable. For example, if we overestimate the order of the VAR, then the singularity of the innovations renders the Yule-Walker equation system singular as well. We are going to show that this has a severe impact on accuracy of predictions. While the asymptotic rate of the mean square prediction error is not impacted by this problem, the finite sample behaviour is severely suffering. This effect will be reinforced, if the predictor variables are not coming from the stationary distribution of the process, but contain additional noise. Again, this happens to be the case in context of dynamic factor models. We will explain the reason for this phenomenon and show how to overcome the problem. Our numerical results underline that it is very important to adapt prediction algorithms accordingly.

Nonparametric causal inference
Organizer: Mats Stensrud
Chair: Mats Stensrud
Room: Pequeno Auditorio

13:30 **A bipartite ranking approach to two-sample nonparametric hypothesis testing**
Myrto Limnios, Stephan Clemencon, Nicolas Vayatis

Abstract: Ranking random observations has become essential to many data analysis problems, ranging from recommendation systems, computational biology, to information retrieval for instance, wherein the information acquisition processes nowadays often involve various and poorly controlled sources, leading to datasets possibly exhibiting strong sampling bias. Fundamental to nonparametric hypothesis testing when the observations are drawn from multiple independent distributions, its study in high-dimension is the subject of much attention, especially due to the lack of natural relation order in the underlying space. In this talk, we will discuss an approach for ranking random observations of complex structure, when drawn from two unknown distributions, relying on a generalization of two-sample linear rank statistics. We will show how this new class encompasses and naturally extends classic univariate rank test statistics, namely for testing homogeneity and statistical independence. We will relate those problems to the concept of Receiver Operating Characteristic (ROC) curve, and provide finite sample guarantees of the testing errors. Convincing experimental studies will illustrate the advantages of this approach compared to state-of-the art methods.

14:00 **A nonparametric Gail-Simon test and estimand for qualitative effect heterogeneity**
Mats Stensrud, Aaron Hudson, Riccardo Brioschi, *Oliver Dukes*

Abstract: Qualitative heterogeneity or effect modification, occur when treatment is beneficial for certain sub-groups and harmful for others. This specific type of heterogeneity is of clinical interest when treatment decisions will be tailored to individual characteristics. The problem of testing for qualitative heterogeneity has been well-studied when the comparison is made between finite subgroups; for example, Gail and Simon (1985) proposed a likelihood ratio test in the context of discrete covariates. However, the problem is more challenging when the potential effect modifiers are continuous, and one wishes to infer the conditional average treatment effect under a nonparametric model. In this talk, I will propose a class of nonparametric tests for qualitative heterogeneity as a natural extension of the Gail-Simon test. Compared with some recent approaches, our proposal can incorporate a variety of structured assumptions on the conditional average treatment effect, extends to moderate/high-dimensional covariates and does not require sample splitting. The utility of the proposal is borne out in simulation studies and a re-analysis of a recent clinical trial.

14:30 **Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space**
Aaron Hudson

Abstract: It is often of interest to assess whether a function-valued statistical parameter, such as a density function or a mean regression function, is equal to any function in a class of candidate null parameters. This can be framed as a statistical inference problem where the target estimand is a scalar measure of dissimilarity between the true function-valued parameter and the closest function among all candidate null values. These estimands are typically defined to be zero when the null holds and positive otherwise. While there is well-established theory and methodology for performing efficient inference when one assumes a parametric model for the function-valued parameter, methods for inference in the nonparametric setting are limited. When the null holds, and the target estimand resides at the boundary of the parameter space, existing nonparametric estimators either achieve a non-standard limiting distribution or a sub-optimal convergence rate, making inference challenging. In this work, we propose a strategy for constructing nonparametric estimators with improved asymptotic performance. Notably, our estimators converge at the parametric rate at the boundary of the parameter space and also achieve a tractable null limiting distribution. As illustrations, we discuss how this framework can be applied to perform inference in nonparametric regression problems, and also to perform nonparametric assessment of stochastic dependence.

15:00 **Kernel Debiased Plug-in Estimation**
Ivana Malenica

Abstract: Modern estimation methods rely on the plug-in principle, which substitutes unknown parameters of the underlying data-generating process with estimated empirical counterparts. Flexible, machine learning (ML) estimation methods have further exploited the plug-in approach. The use of highly adaptive, complex ML algorithms, however, induces plug-in bias (first-order bias) that impacts the downstream estimate. Traditional methods addressing this sub-optimal bias-variance trade-off rely on the efficient influence function (EIF) of the target parameter. When estimating multiple target parameters, these methods require debiasing the nuisance parameter multiple times using the corresponding EIFs, posing analytical and computational challenges. In this work, we leverage the targeted maximum likelihood estimation framework to propose a novel method named kernel debiased plug-in estimation (KDPE). KDPE refines an initial estimate through regularized likelihood maximization steps, employing a nonparametric model based on reproducing kernel Hilbert spaces. We show that KDPE (i) simultaneously debiases all pathwise differentiable target parameters that satisfy our regularity conditions, (ii) does not require the EIF for implementation, and (iii) remains computationally tractable. We numerically illustrate the use of KDPE and validate our theoretical results.

Recent advances in semiparametric and nonparametric econometrics
Organizer: Juan Carlos Escanciano
Chair: Juan Carlos Escanciano
Room: Grande Auditorio

13:30 **Regular identification of the ATE without the strict overlap condition**
Telmo Pérez-Izquierdo

Abstract: Estimation of the Average Treatment Effect (ATE) in the selection on observables framework usually relies on the strict overlap assumption: that the propensity score is bounded away from zero and one. In this paper, I derive a milder necessary and sufficient condition for regular identification (positive Semiparametric Fisher Information) of the ATE. The result is based on a generalization of van der Vaart (1991)'s result linking differentiability and Fisher Information. The condition for the regular identification of the ATE is based on a square-integrability condition involving the conditional variances of potential outcomes and the propensity score. This highlights a trade-off between assumptions on the distribution of potential outcomes given observed covariates and assumptions on selection onto treatment. I discuss the implications of the milder condition for empirical research and its distinction from the strict overlap assumption. For example, I demonstrate that the ATE is regularly identified in the case of a single normal regressor and logistic error on the selection equation, even if the strict overlap condition is violated.

14:00 **On the Existence and Information of Orthogonal Moments**
Juan Carlos Escanciano, Facundo Argañaraz

Abstract: Locally Robust (LR)/Orthogonal/Debiased moments have proven useful with machine learning first steps, but their existence has not been investigated for general parameters. In this paper, we provide a necessary and sufficient condition, referred to as Restricted Local Non-surjectivity (RLN), for the existence of such orthogonal moments to conduct robust inference on general parameters of interest in regular semiparametric models. Importantly, RLN does not require either identification of the parameters of interest or the nuisance parameters. However, for orthogonal moments to be informative, the efficient Fisher Information matrix for the parameter must be non-zero (though possibly singular). Thus, orthogonal moments exist and are informative under more general conditions than previously recognized. We demonstrate the utility of our general results by characterizing orthogonal moments in a class of models with Unobserved Heterogeneity (UH). For this class of models our method delivers functional differencing as a special case. Orthogonality for general smooth functionals of the distribution of UH is also characterized. As a second major application, we investigate the existence of orthogonal moments and their relevance for models defined by moment restrictions with possibly different conditioning variables. We find orthogonal moments for the fully saturated two stage least squares, for heterogeneous parameters in treatment effects, for sample selection models, and for popular models of demand for differentiated products. We apply our results to the Oregon Health Experiment to study heterogeneous treatment effects of Medicaid on different health outcomes.

14:30 **Estimation and inference of panel data models with a generalized factor structure**
Juan Manuel Rodriguez-Poo, Alexandra Soberon, Stefan Sperlich

Abstract: It has become increasingly popular to apply not only fixed effects but also unobserved factors to control in panel data models for unobserved heterogeneity that could cause omitted variable biases. Besides many advantages, we see in the present literature basically three flaws: the unjustified and unnecessary strong parametric restrictions on the nuisance part in those models, the lack of interpretation for maybe most of the variation in the data, and the absence of specification tests that checks the hypotheses made for the pursued approach. This article (re-)introduces therefore a modeling that attacks the second flaw, using nonparametric methods to circumvent the first flaw, and derives a specification test to verify the potentially most critical assumption made in our modeling. For the parameters that typically attract most of the attention, we derive consistent estimators that are asymptotically normal at rate root NT, while for the nuisance part we obtain the optimal nonparametric one. A nonparametric specification test checks the crucial modeling assumption. It relies on combining the methodology of conditional moment tests and nonparametric estimation techniques. Using degenerate and nondegenerate theories of U-statistics we can show that its convergence and asymptotically distribution under the null, and that it diverges under the alternative at a rate arbitrarily close to sqrt-NT. Finite sample inference is based on bootstrap. We evaluate the performance of our methods by means of simulations, followed by an empirical study of the effect of EU ETS on the economic development of EU countries.

15:00 **Chi-square goodness-of-fit tests to check for conditional moment restrictions**
Miguel A. Delgado, Antonio Raiola

Abstract: This paper proposes chi-square type tests for assessing the specification of regression models or general conditional moment restrictions. The data is partitioned according to the explanatory variables into several cells, and the tests evaluate whether the difference between the observed average of the dependent variable and its expected value under the model specification, in each cell, arises by chance. In contrast to existing omnibus procedures, the proposed tests are asymptotically pivotal and fairly insensitive to the curse of dimensionality. The computation is straightforward and can be implemented without the assistance of bootstrap techniques. Importantly, the asymptotic properties of the test are invariant to sample-dependent partitions, which can be chosen to favor certain alternatives. A Monte Carlo study provides evidence of the performance of the tests using samples of fairly small size compared with existing omnibus alternatives, particularly when there are many explanatory variables. An empirical application regarding returns to education of African American students in the US complements the finite sample study.

15:30 - 16:00 Coffee Break

16:00 - 17:00 ISNPS General Assembly
Room: Grande Auditorio

17:00 - 19:00 Invited 4

Modern advances at the interface of statistical learning and inference
Organizer: Pragya Sur
Chair: Subhabrata Sen
Room: Grande Auditorio

17:00 **Minimax estimation in Efron's two-groups model**
Chao Gao

Abstract: The advent of large scale inference has spurred reexamination of conventional statistical thinking. In a series of highly original articles Efron showed in some examples that the ensemble of the null distributed test statistics grossly deviated from the theoretical null distribution and Efron persuasively illustrated the danger in assuming the theoretical null's veracity for downstream inference. Though intimidating in other contexts the large scale setting is to the statistician's benefit here. There is now potential to estimate rather than assume the null distribution. In a model for n many z-scores with at most k nonnulls we adopt Efron's suggestion and consider estimation of location and scale parameters for a Gaussian null distribution. Placing no assumptions on the nonnull effects we consider rate-optimal estimation in the entire regime $k < n/2$ that is precisely the regime in which the null parameters are identifiable. The minimax upper bound is obtained by considering estimators based on the empirical characteristic function and the classical kernel mode estimator. Faster rates than those in Huber's contamination model are achievable by exploiting the Gaussian character of the data. As a consequence it is shown that consistent estimation is indeed possible in the practically relevant regime $k \asymp n$. In a certain regime the minimax lower bound involves constructing two marginal distributions whose characteristic functions match on a wide interval containing zero. The construction notably differs from those in the literature by sharply capturing a second-order scaling of $n/2 - k$ in the minimax rate.

17:30 **Denoising over network with application to partially observed epidemics**
Olga Klopp, Claire Donnat, Nicolas Verzelen

Abstract: We introduce a novel approach to predict epidemic spread over networks using total variation (TV) denoising, a signal processing technique. The study proves the consistency of TV denoising with Bernoulli noise, extending existing bounds from Gaussian noise literature. The methodology is further extended to handle incomplete observations, showcasing its effectiveness. We show that application of 1-bit total variation denoiser improves the prediction accuracy of virus spread dynamics on networks.

18:00 **Fast Linear Model Trees by PILOT**
Peter Rousseeuw, Jakob Raymaekers, Tim Verdonck, Ruicong Yao

Abstract: Linear model trees are regression trees that incorporate linear models in the leaf nodes. This preserves the intuitive interpretation of decision trees and at the same time enables them to better capture linear relationships, which is hard for standard decision trees. But most existing methods for fitting linear model trees are time consuming and therefore not scalable to large data sets. In addition, they are more prone to overfitting and extrapolation issues than standard regression trees. In this paper we introduce PILOT, a new algorithm for linear model trees that is fast, regularized, stable and interpretable. PILOT trains in a greedy fashion like classic regression trees, but incorporates an L_2 boosting approach and a model selection rule for fitting linear models in the nodes. The abbreviation PILOT stands for PIecewise Linear Organic Tree, where `organic' refers to the fact that no pruning is carried out. PILOT has the same low time and space complexity as CART without its pruning. An empirical study indicates that PILOT tends to outperform standard decision trees and other linear model trees on a variety of data sets. Moreover, we prove its consistency in an additive model setting under weak assumptions. When the data is generated by a linear model, the convergence rate is faster.

18:30 **Adaptive Inference in Sequential Experiments**
Cun-Hui Zhang, Mufang Ying, Koulik Khamaru

Abstract: Sequential data collection has emerged as a widely adopted technique for enhancing the efficiency of data gathering processes. Despite its advantages, such data collection mechanism often introduces complexities to the statistical inference procedure. For instance, the ordinary least squares estimator in an adaptive linear regression model can exhibit non-normal asymptotic behavior, posing challenges for accurate inference and interpretation. We propose a general method for constructing debiased estimator which remedies this issue. The idea is to make use of adaptive linear estimating equations. We establish theoretical guarantees of asymptotic normality, supplemented by discussions on achieving near-optimal asymptotic variance.

Causal inference for studying vaccine effects
Organizer: Daniel Nevo
Chair: Aaron Hudson
Room: Pequeno Auditorio

17:00 **Evaluating immune correlates of protection in vaccine efficacy trials with stochastic-interventional causal effects**
Nima Hejazi

Abstract: In vaccine efficacy clinical trials randomizing participants to active v control conditions and following individuals until the occurrence of an infectious disease outcome of interest, evaluating the efficacy of a candidate vaccine through immune markers specified a priori is complicated by factors including (1) the relative dearth of causal inference approaches tailored to quantitative mediators and (2) the challenge of correcting for outcome-dependent (e.g., case-cohort) sampling used to measure such markers. We present a two-pronged solution, using (1) modified treatment policies to formulate interventions on immune markers whose identifiability can be rigorously probed and (2) inverse probability of sampling weights to obtain population-level inference. We outline non-/semi-parametric inference strategies that yield asymptotically efficient estimators of our proposed causal estimands and that incorporate machine learning. Our effect definitions measure the causal effect of perturbing an immune marker’s observed value in the active condition, resulting in an interpretable causal dose-response analysis informing on potential vaccine protection mechanisms and aiding in identification of surrogate endpoints. We demonstrate their utility by highlighting applications to modeling how modifications to a vaccine tested in a phase 3 trial would be expected to alter vaccine efficacy across four distinct trials of the Coronavirus Prevention Network and HIV Vaccine Trials Network.

17:30 **Nonparametric Identification of Immunologic and Behavioral Effects in Vaccination Studies**
Daniel Nevo, Mats Stensrud, Uri Obolski

Abstract: In randomized trials assessing vaccination effects, the interpretation of trial results and vaccine efficacy estimands is subtle. In this talk, we introduce a framework to define different vaccine efficacy estimands and clarify their interpretations. This framework allows us to explicitly consider immunologic and behavioral effects of vaccination. We highlight how policy-relevant estimands may substantially deviate from those traditionally reported in vaccine studies. We present plausible technical conditions under which a conventional vaccine trial allows identification and estimation of different vaccine estimands. Central to our approach is the introduction of a post-treatment variable, a ``belief variable'', that indicates the treatment an individual believed they had received. We also present an alternative identification assumptions under weaker assumptions involving the additional measurement of adverse effects. We discuss estimation and illustrate the relations between the different estimands, and their practical relevance, in numerical examples based on an influenza vaccine trial.

18:00 **Waning of treatment effects**
Mats Stensrud, Matias Janvin

Abstract: Knowing whether a treatment effect declines over time is important for health policy and drug development. A classical example is the waning of vaccine effects. However, defining and choosing an adequate measure of waning is challenging. One strategy is to record antibody levels in immunological assays after vaccination, but the level of antibodies does not fully capture susceptibility towards infectious outcomes. Another strategy is to contrast vaccine efficacy at different times after treatment administration in a randomized experiment, but a simple contrast of vaccine efficacy at two different times compares different populations of individuals: those who were uninfected at the first time versus those who remain uninfected until the second time. Thus, the contrast of vaccine efficacy at early and late times can not be interpreted as a causal effect. Here we introduce a different estimand, the challenge effect, which corresponds to the hypothetical efficacy that would be observed under an intervention that isolates every individual until the time of interest, and then exposes every individual to the infectious agent. We identify sharp bounds on the challenge effect under plausible, non-parametric assumptions that are broadly applicable in studies of waning, in particular in vaccine trials using routinely collected data. Finally, we apply the methods to derive informative bounds on the waning of the BNT162b2 COVID-19 vaccine.

18:30 **Vaccine effectiveness estimation under the test-negative design: identifiability and efficiency theory for causal inference under conditional and control exchangeability**
Cong Jiang, Denis Talbot, Sara Carazo, *Mireille Schnitzer*

Abstract: The test-negative design (TND) is routinely used for the monitoring of seasonal flu vaccine effectiveness and recently become integral to COVID-19 vaccine surveillance, notably in Québec, Canada. Distinct from the case-control study, it involves the recruitment of participants with a common symptom presentation who are then tested for the target infection. Participants with positive tests are considered "cases," while those with negative tests are "controls." Logistic regression has traditionally been used to adjust for confounders to estimate vaccine effectiveness under the TND. However, this approach may be biased if effect modification by a confounder exists or if the model is otherwise misspecified. We first review an inverse probability of treatment weighting estimator for the marginal risk ratio that is valid under effect modification but requires parametric modeling for the conditional probability of vaccination. To address this limitation, we propose a novel doubly robust and efficient estimator of the marginal risk ratio. We theoretically and empirically demonstrate the parametric convergence rates achieved through machine learning of the nuisance functions. Our study was motivated by the goal of improving adjustment for measured confounders when estimating COVID-19 vaccine effectiveness among community-dwelling people 60 years and older in Québec, and we provide results for this application.

Computer-intensive methods for complex data

Organizer: Dimitris Politis
Chair: Dimitris Politis
Room: Sala Polivalente 1.1

17:00 **Comparing many functional means**
Stanislav Volgushev, Dehan Kong, Colin Decker

Abstract: Many modern medical devices produce data with the structure of a multi-channel functional time series. Examples include medical imaging devices (fMRI, EEG, and ECG), high through-put time course gene sequencing devices, and high through-put devices that measure time-course microbiome composition. The typical number of channels can be of the same order or larger than the available sample size. In this talk, we will present methodology to simultaneously test the equality of a growing number of functional means, in the example above, each mean corresponds to a channel. The number of channels can grow exponentially in the sample size. The proposed test is fully functional in the sense that we do not conduct any explicit dimension reduction or principal component analysis. The practical implementation is based on a Gaussian multiplier procedure and we provide explicit bounds on the speed of convergence of the rejection probability of our test to the nominal value under the null and power against local alternatives. Our theoretical analysis leverages recent advances in high-dimensional Gaussian approximation but requires several intricate modifications of those techniques.

17:30 **Statistical inference with optimal sampling**
Alan Welsh, *Nan Zou*

Abstract: In classic statistical inference, Ordinary Least Squares regression (OLS) has been the workhorse in studying the effect of one or more predictors on response. However, for datasets with massive sample sizes, which are increasingly prevalent these days, the OLS can be computationally infeasible. To speed up the OLS for massive datasets, the optimal sampling OLS selects an optimal small subset of samples from the original massive number of samples. Despite its considerable popularity, it seems unclear what conditions can guarantee optimal sampling OLS's asymptotical normality. This talk will first introduce the optimal sampling OLS procedure and then investigate the conditions for its asymptotical normality. Specifically, it seems (1) when the number of predictors is fixed, the optimal sampling OLS is asymptotically normal if and only if the OLS itself is asymptotically normal, and (2) when the number of predictors goes to infinity, the optimal sampling OLS requires a more restrictive condition than the OLS.

18:00 **Bootstrapping "Likehihood Ratio tests" under mispecification**
Pascal Lavergne, *Patrice Bertail*

Abstract: We consider likelihood ratio tests for restrictions on parameters, where likelihood ratio broadly refers to any statistic built from a convex M-estimation criterion, including quantile regression. We consider potential model misspecification, which renders the likelihood ratio statistic not asymptotically pivotal. We propose a general and simple nonparametric bootstrap procedure that yields asymptotically valid critical values. The method modifies the bootstrap objective function to mimic what happens under the null hypothesis. Extensions to the Markovian cases are discussed. A Monte-Carlo study illustrates that a double bootstrap likelihood ratio test controls level well and is powerful.

18:30 **Bootstrap-assisted inference for weakly stationary time series**
Yunyi Zhang

Abstract: The literature often adopts two types of stationarity assumptions in the analysis of time series, i.e., the weak stationarity, suggesting that the mean and the autocovariance function of a time series are time invariant; and strict stationarity, indicating that the marginal distributions of the time series are time invariant. While the strict stationarity assumption is vital from theoretical aspect, it is hard to verify in practice. On the other hand, the weak stationarity is relatively feasible to ensure and verify, as it only relies on the second--order structures of the time series. Concerning this, while sorts of weak stationarity assumptions are typically adopted in time series modeling, statisticians may want to avoid relying on strict stationarity assumptions during statistical inference. This presentation focuses on the analysis of quadratic forms within a weakly, but not necessarily strictly stationary (vector) time series. In the context of scalar time series, it establishes the Gaussian approximation for quadratic forms of a short--range dependent weakly stationary scalar time series. Building upon this result, it derives the asymptotic distributions of the sample autocovariances, the sample autocorrelations, and the sample autoregressive coefficients. Transitioning to vector time series, this presentation tackles statistical inference within high--dimensional vector autoregressive models with white noise innovations. Given the complicated covariance structures inherent in non-stationary time series, this presentation adopts the dependent wild bootstrap method to facilitate statistical inference. Numerical results verifies the consistency of the proposed theories and methods. Strict stationarity is hard to ensure and verify for a real--life dataset. Therefore, our work should be able to assist statisticians in capturing the inherent non--stationarity of real--life time series.

Topics in Econometrics: Big Data, Panel Estimation, and Forecasted Treatment

Organizer: Jeffrey Racine
Chair: Jeffrey Racine
Room: Sala Polivalente 1.2

17:00 **A Robust Method for Microforecasting and Estimation of Random Effects**
Silvia Sarpietro, Raffaella Giacomini, Sokbae Lee

Abstract: We propose a method for forecasting individual outcomes and estimating random effects in linear panel data models and value-added models when the panel has a short time dimension. The method is robust, trivial to implement and requires minimal assumptions. The idea is to take a weighted average of time series- and pooled forecasts/estimators, with individual weights that are based on time series information. We show the forecast optimality of individual weights, both in terms of minimax-regret and of mean squared forecast error. We then provide feasible weights that ensure good performance under weaker assumptions than those required by existing approaches. Unlike existing shrinkage methods, our approach borrows the strength - but avoids the tyranny - of the majority, by targeting individual (instead of group) accuracy and letting the data decide how much strength each individual should borrow. Unlike existing empirical Bayesian methods, our frequentist approach requires no distributional assumptions, and, in fact, it is particularly advantageous in the presence of features such as heavy tails that would make a fully nonparametric procedure problematic.

17:30 **Partial identification in nonlinear panels**
Chris Muris

Abstract: We propose a systematic approach to characterizing the identified set for common parameters and partial effects for a large class of nonlinear panel models with fixed effects. Our method applies generally to nonlinear panel models with a likelihood representation, and it generates new results for a number of semiparametric static and dynamic discrete choice models. Computation of the bounds is simple and fast. Our identification results suggest novel estimators for the identified set.

18:00 **Fast Inference for Quantile Regression with Tens of Millions of Observations**
Youngki Shin, Yuan Liao, Myung Hwan Seo, Sokbae Lee

Abstract: Big data analytics has opened new avenues in economic research, but the challenge of analyzing datasets with tens of millions of observations is substantial. Conventional econometric methods based on extreme estimators require large amounts of computing resources and memory, which are often not readily available. In this paper, we focus on linear quantile regression applied to "ultra-large" datasets, such as U.S. decennial censuses. A fast inference framework is presented, utilizing stochastic subgradient descent (S-subGD) updates. The inference procedure handles cross-sectional data sequentially: (i) updating the parameter estimate with each incoming "new observation",(ii) aggregating it as a Polyak-Ruppert average, and (iii) computing a pivotal statistic for inference using only a solution path. The methodology draws from time-series regression to create an asymptotically pivotal statistic through random scaling. Our proposed test statistic is calculated in a fully online fashion and critical values are calculated without resampling. We conduct extensive numerical studies to showcase the computational merits of our proposed inference. For inference problems as large as (n,d) ~ (10^7, 10^3), where n is the sample size and d is the number of regressors, our method generates new insights, surpassing current inference methods in computation. Our method specifically reveals trends in the gender gap in the U.S. college wage premium using millions of observations, while controlling over 10^3 covariates to mitigate confounding effects.

18:30 **Forecasted Treatment Effects**
Irene Botosaru

Abstract: We consider estimation and inference of the effects of a policy in the absence of a control group. We obtain unbiased estimators of individual (heterogeneous) treatment effects and a consistent and asymptotically normal estimator of the average treatment effects, based on forecasting counterfactuals using a short time series of pre-treatment data. We show that the focus should be on forecast unbiasedness rather than accuracy. Correct specification of the forecasting model is not necessary to obtain unbiased estimates of individual treatment effects. Instead, simple basis function (e.g., polynomial time trends) regressions deliver unbiasedness under a broad class of data-generating processes for the individual counterfactuals. Basing the forecasts on a model can introduce misspecification bias and does not necessarily improve performance even under correct specification. Consistency and asymptotic normality of our Forecasted Average Treatment effects (FAT) estimator are attained under an additional assumption that rules out common and unforecastable shocks occurring between the treatment date and the date at which the effect is calculated.

Statistics for spatial and network data

Organizer: Soumendra Lahiri
Chair: Soumendra Lahiri
Room: Sala Polivalente 1.3

17:00 **Variance Estimation of Spectral Statistics for Spatial Processes using Subsampling**
Souvick Bera, Daniel Nordman, *Soutir Bandyopadhyay*

Abstract: In the realm of frequency domain analysis for spatial data, estimators based on the periodogram often exhibit complex variance structures originating from aggregated periodogram covariances. Previous attempts to bootstrap these statistics face challenges in capturing these variances and quantifying estimation uncertainty. This difficulty arises because achieving consistency for various periodogram-based statistics requires evaluating the periodogram at an increasing number of frequencies as the sample size grows. Despite the diminishing dependence between periodogram ordinates, the decay rate balances the growing frequencies, preserving a dependence structure in the limiting distribution. Consequently, the validity of frequency domain bootstrap (FDB) approaches for spatial data is confined to a specific class of processes and statistics. To overcome this challenge, we propose subsampling and cutting-edge FDB methods based on subsampling. These methods can accurately capture uncertainty without necessitating additional stringent assumptions beyond those required for the existence of a target limit distribution of spectral statistics - a departure from typical bootstrap practices. Moreover, our work fills a gap in the theory of subsampling for spatial data by providing distributional approximations, alongside variance estimation, for spectral statistics.

17:30 **Empirical likelihood inference in the frequency domain for dependent data**
Dan Nordman, Haihan Yu, Mark Kaiser

Abstract: Frequency domain analysis of time series is often complicated by periodogram-based statistics having complex variances, so that approximations from resampling or empirical likelihood (EL) can be helpful. Existing versions of periodogram-based EL for time series, though, are restricted in applicability; these are not valid outside of special linear processes and spectral parameters. This talk describe a new spectral EL (SEL) method by merging two different EL frameworks for time series, namely, block-based and periodogram-based EL. The resulting SEL statistics have some nice features for inference: these admit chi-square limits under mild conditions and can be coupled to an effective bootstrap procedure. The scope of EL for time series inference is then greatly expanded as SEL: can handle any spectral parameters; is valid for general processes (including nonlinear); and has a provable bootstrap that provides a novel alternative to other resampling plans in the frequency domain. Numerical studies suggest that the method has good accuracy performance, and the approach is also demonstrated with a real example.

18:00 **Graph wavelet variances**
Debashis Mondal, Rodney Fonseca, Aluisio Pinheiro

Abstract: In this talk, we introduce a new concept called graph wavelet variance to analyze the variability of observations on irregular graphs. The proposed graph wavelet variance allows for an evaluation of variability on different scales of the graph spectrum, and offers valuable insights on the stochastic behavior and the dependence structure. We present estimation of the graph wavelet variance, derive asymptotic results under appropriate dependence structure, and discuss ways to compute approximate confidence intervals. Unlike the graph Fourier transform, wavelet transforms on irregular graphs are computed without the full spectral decomposition of the Laplacian matrix, thereby making the proposed idea particularly useful when observations are made on irregular graph. The proposed method is applied to analyze spatial-temporal dynamics of Sars-Cov-2 infection rates in Brazil, using cities/ population centers as graph nodes and shared borders as edges or links. Through the analysis of graph wavelet variances, we characterizes the spread of the pandemic at three different time periods in the country.

18:30 **Conformal Prediction for Network-Assisted Regression**
Robert Lunde, Elizaveta Levina, Ji Zhu

Abstract: An important problem in network analysis is predicting a node attribute using nodal covariates and summary statistics computed from the network, such as graph embeddings or local subgraph counts. While standard regression methods may be used for prediction, statistical inference is complicated by the fact that the nodal summary statistics often exhibit a nonstandard dependence structure. We show that conformal prediction offers finite-sample valid prediction intervals for network-assisted regression under a joint exchangeability condition and a mild regularity condition on the network statistics. In the case where the observed graph is a non-representative sample of the population, we show that conformal prediction remains finite-sample valid for a randomly chosen test point from the sample if the sampling mechanism satisfies an invariance property. We show that this invariance condition is satisfied for selection rules related to widely used sampling schemes such as ego sampling and snowball sampling.

Topics in nonparametric and semiparametric econometrics

Organizers: Hira Koul and Indeewara Perera
Chair: Indeewara Perera
Room: Sala Polivalente 1.4

17:00 **Estimation of Grouped Time-Varying Network Vector Autoregressive Models**
Degui Li, Bin Peng, Songqiao Tang, Weibiao Wu

Abstract: This paper introduces a flexible time-varying network vector autoregressive model framework for large-scale time series. A latent group structure is imposed on the heterogeneous and node-specific time-varying momentum and network spillover effects so that the number of unknown time-varying coefficients to be estimated can be reduced considerably. A classic agglomerative clustering algorithm with nonparametrically estimated distance matrix is combined with a ratio criterion to consistently estimate the latent group number and membership. A post-grouping local linear smoothing method is proposed to estimate the group-specific time-varying momentum and network effects, substantially improving the convergence rates of the preliminary estimates which ignore the latent structure. We further modify the methodology and theory to allow for structural breaks in either the group membership, group number or group-specific coefficient functions. Numerical studies including Monte-Carlo simulation and an empirical application are presented to examine the finite-sample performance of the developed model and methodology.

17:30 **Inference of Unknown Semiparametric Transformation via Distribution Regression Estimation**
Yi He, Juan-Juan Cai

Abstract: We introduce a novel method for estimating and inferring transformed linear regression models with an unknown transformation of the dependent variable. By reformulating the regression as a distribution regression model, we directly estimate and infer the unknown transformation as the intercept function. We establish the joint asymptotic normality of the transformation function and linear regression coefficient estimators. In contrast to existing theory, we consider a more flexible setting with fixed-design, i.e., non-stochastic regressors. We validate the effectiveness of bootstrap methods. Finally, we apply our method to analyze a p-hacking dataset in economics.

18:00 **Bootstrap specification tests for multivariate GARCH processes**
Indeewara Perera, Kanchana Nadarajah

Abstract: We develop tests for the correct specification of the conditional distribution in multivariate GARCH models based on empirical processes. We transform the multivariate data into univariate data based on the marginal and conditional cumulative distribution functions specified by the null model. The test statistics considered are based on empirical processes of the transformed data in the presence of estimated parameters. The limiting distributions of the proposed test statistics are model dependent and are not free from the underlying nuisance parameters, making the tests difficult to implement. To address this, we develop a novel bootstrap procedure which is shown to be asymptotically valid irrespective of the presence of nuisance parameters. This approach utilises a particular scalable iterated bootstrap method and is simple to implement as the associated test statistics have simple closed form expressions. A simulation study demonstrates that the new tests perform well in finite samples. A real data example illustrates the testing procedure.

18:30 **Regression Modelling under General Heterogeneity**
Liudas Giraitis, Yufei Li, George Kapetanios

Abstract: This paper outlines the unrestrictive environment permitting general heterogeneity in regression modelling. It shows that regression models with fixed and time-varying parameters remain meaningful and can be estimated by the OLS and time-varying OLS methods for a very wide class of regressors and regression noises which use is not permitted in the standard regression modelling. It describes a wide class of regressors which allows to develop an asymptotic theory, to estimate standard errors and to construct confidence intervals for parameters. It shows that estimation of robust confidence intervals permits a very high degree of heterogeneity in regressors and regression noise, and differently from their theoretical asymptotic counterparts, their estimation is easy. The estimator of robust standard errors is very similar to the well-known estimator of heteroskedasticity-consistent standard errors by White (1980). The robust standard errors are easy to compute, and the robust confidence intervals perform well in Monte Carlo simulations. This makes them attractive in applied work. The paper includes an empirical experiment.

Semi- and non-parametric approaches for inference on high dimensional data

Organizer: Wen Zhou
Chair: Wen Zhou
Room: Sala Polivalente 1.5

17:00 **Innovative unsupervised approach for simultaneous subgroup recovery and group-specific feature identification**
Lyuou Zhang, Xiwei Tang, *Wen Zhou*

Abstract: Simultaneously identifying heterogeneous subgroups and the informative features defining them, especially in the absence of responses and with a plethora of features, has long been a challenge in various domains, including omics studies, clinical research, and policy evaluation. Existing methods have either focused narrowly on global informative features or performed feature selection and group recovery as separate tasks, overlooking their interactions. Such methods might miss scientifically relevant information, and lead to suboptimal solutions to both feature identification and subgroup recovery. To overcome these limitations, we introduce a novel unsupervised learning approach, PAirwise REciprocal fuSE (PARSE), which concurrently pinpoints cluster-specific informative features and conducts high-dimensional clustering. Our method employs a new regularization that heavily penalizes features with minor differences across clusters, thus avoiding the selection of less informative features that define clusters. The oracle property of PARSE is obtained, and we establish lower bounds for both clustering and cluster-specific feature identification, affirming our method's optimality in both aspects. For implementations, we have devised an enhanced Expectation-Maximization algorithm, which is computationally feasible. Extensive numerical studies showcase PARSE's superiority over existing methods. In an application involving the identification of gene signatures in different subtypes of human pancreatic cells using single-cell RNAseq data, PARSE outperforms most mainstream methods in terms of identifying both the cell subtypes and corresponding gene signatures.

17:30 **Multidimensional Signal-to-Noise Ratio Estimation for High Dimensional Random Effects Models under Heteroscedasticity**
Xiaodong Li, Xiaohan Hu, Zhentao Li

Abstract: Variance and signal-to-noise estimation with high dimensional random effects models has recently been widely used in genomics for the purpose of model-based heritability estimation, and extensions to multivariate traits also attract much attention in various phenotypically rich studies. In this talk, I will introduce our recent results on estimation and inference about these parameter under high-dimensional random effects models with multivariate response and heterogeneous noise. We consider two methods: method of moments and a likelihood based aggregated estimating equation method. For each method, we establish the consistency and asymptotic distribution of the estimator, in particular on how the asymptotic standard errors rely on the noise heteroskedasticity. Extensive numerical experiments illustrate our theoretical findings. This is a joint work with my PhD students Xiaohan Hu and Zhentao Li.

18:00 **Local perspectives in latent space network models**
Lijia Wang, YX Rachel Wang, Xin Tong, Xiao Han, Yanhui Wu

Abstract: The impact of neighborhood dynamics in social networks is a critical factor in shaping an individual's decision-making and opinion formation. To comprehend how social networks influence individual behaviors and viewpoints, it's essential to analyze an individual's localized perspective against the backdrop of the broader network. In this study, we explore a general latent space network model and address the challenge of deducing the latent positions of nodes using only the partial information available within an individual's localized information. By implementing a projected gradient descent approach, we demonstrate that the rate of convergence for our estimates is influenced by the neighborhood characteristics of the individual node. We introduce a metric to quantify the level of bias present in an individual's localized perspective. Through the application of our methods to both real and simulated networks, including the cosponsorship network in the US Congress, we contrast these local latent position estimates against global ones. Our findings illustrate how our theoretical framework enhances our grasp of the intricate local viewpoints that exist within social networks.

18:30 **Sparse Heteroskedastic PCA in High Dimensions**
Zhao Ren, Rui Kang, Peiliang Zhang

Abstract: Principal Component Analysis (PCA) is a widely adopted multivariate statistical technique, renowned for its versatility across numerous disciplines. To address the challenges posed by high-dimensional and heteroskedastic data, we consider a general framework in high dimensions, the generalized spiked covariance model with sparse loadings. We propose a novel algorithm called \emph{SparseHPCA} that leverages the orthogonal iteration method, noise-adaptive thresholding, and diagonal imputation techniques. The proposed procedure is computationally feasible and fully data-driven without prior knowledge about the noise levels and the sparsity of the loading matrix. Theoretical analysis shows that SparseHPCA achieves minimax optimal convergence rates. By applying SparseHPCA, we further investigate the Sparse SVD problem in the presence of heteroskedastic noise. The effectiveness of the proposed method is revealed through experiments on both simulated and two real datasets.

9:00 - 11:00 Invited 5

Nonparametric estimation in high dimensions
Organizer: Zhou Fan
Chair: Nikolaos Ignatiadis
Room: Grande Auditorio

9:00 **Rate Optimality and Phase Transition for User-Level Local Differential Privacy**
Yi Yu

Abstract: Most of the literature on differential privacy considers the item-level case where each user has a single observation, but a growing field of interest is that of user-level privacy where each user holds multiple observations and wishes to maintain the privacy of their entire collection. In this paper, we prove a general minimax lower bound, which shows that, for any locally private user-level estimation problem, the risk cannot be made to vanish for a fixed number of users even when each user holds an arbitrarily large number of observations. We then prove tight minimax lower and upper bounds for univariate and multidimensional mean estimation, sparse mean estimation, and non-parametric density estimation. In particular, we observe a phase-transition in the rate when the number of samples each user holds is sufficiently large relative to the number of users. Further, in the case of (non-sparse) mean estimation and density estimation, we see that, up until the phase transition, the rate is the same as having an equivalent number of users in the item-level setting, matching similar behaviour seen in other user-level results from previous works. However different behaviour is observed in the case of sparse mean estimation wherein this problem is infeasible when the dimension exceeds the number of observations in the item-level setting, but is tractable in the user-level setting. The estimator recovers elements of the performance of the non-private estimator, which may be of independent interest for applications as an example of a high-dimensional problem that is feasible under local privacy constraints.

9:30 **Fundamental limits of community detection from multi-view data**
Subhabrata Sen, Xiaodong Yang, Buyu Lin

Abstract: Multi-view data arises frequently in modern network analysis e.g. relations of multiple types among individuals in social network analysis, longitudinal measurements of interactions among observational units, annotated networks with noisy partial labeling of vertices etc. We will discuss community detection in these disparate settings via a unified framework. In particular, we characterize the sharp thresholds for weak recovery in the inhomogeneous multi-layer stochastic block model and the dynamical stochastic block model. We will also discuss community recovery algorithms based on Approximate Message Passing. Based on joint work with Xiaodong Yang and Buyu Lin (Harvard).

10:00 **Spectrum-Aware Debiasing: A Modern Inference Paradigm with Applications to Principal Component Regression**
Pragya Sur

Abstract: Debiasing methodologies have emerged as a solid toolbox for producing inference in high-dimensional problems. Since its original introduction, the methodology witnessed a major upheaval with the introduction of debiasing with degrees-of-freedom adjustment. Despite overcoming limitations with initial debiasing techniques, this updated method suffers a key issue—the method relies on Gaussian designs and independent, identically distributed samples. In this talk, we will break this barrier via a novel debiasing formula that exploits the spectrum of the sample covariance matrix. Our formula applies for a broad class of non-Gaussian designs, including some heavy-tailed ones, as well as certain dependent data settings. Our correction term differs significantly from the one proposed in prior work but recovers the Gaussian formula as a special case. We will demonstrate the utility of our approach with regard to several statistical inference problems. As a by-product, our work also introduces the first debiased principal component regression estimator with formal guarantees in high dimensions. This is based on joint work with Yufan Li.

10:30 **Tightness of SDP and Burer-Monteiro Factorization for Phase Synchronization in High Noise Regime**
Anderson Ye Zhang

Abstract: We study the phase synchronization problem in the presence of Gaussian noise. For this problem, the maximum likelihood estimation (MLE) is computationally challenging, and hence is relaxed to a semi-definite programming (SDP). When the noise is small, existing literature shows that the solution of SDP is exactly equal to the MLE. However, it remains unclear what happens when the noise is large. In this work, we investigate the distance between the SDP solution and the MLE, that is, quantifying the tightness of the SDP relaxation compared to the MLE. We establish that, in the high-noise regime, the distance between them is exponentially small, with the signal-to-noise ratio appearing in the exponent. By increasing the signal-to-noise ratio, our result shows they coincide with each other, recovering the existing results in the low-noise regime. We further extend our analysis to the Burer-Monteiro factorization of the SDP and establish similar results.

Advanced inference of complex data

Organizer: Regina Liu
Chair: Dimitris Politis
Room: Pequeno Auditorio

9:00 **Fair conformal prediction and risk control**
Linjun Zhang

Abstract: Multi-calibration is a powerful and evolving concept originating in the field of algorithmic fairness. For a predictor f(x) that estimates the outcome y given covariates x, and for a function class C, multi-calibration requires that the predictor f(x) and outcome y are indistinguishable under the class of auditors in C. Fairness is captured by incorporating demographic subgroups into the class of functions C. Recent work has shown that, by enriching the class C to incorporate appropriate propensity re-weighting functions, multi-calibration also yields target-independent learning, wherein a model trained on a source domain performs well on unseen, future target domains(approximately) captured by the re-weightings. The multi-calibration notion is extended, and the power of an enriched class of mappings is explored. HappyMap, a generalization of multi-calibration, is proposed, which yields a wide range of new applications, including a new fairness notion for uncertainty quantification (conformal prediction), a novel technique for conformal prediction under covariate shift, and a different approach for fair risk control, while also yielding a unified understanding of several existing seemingly disparate algorithmic fairness notions and target-independent learning approaches. A single HappyMap meta-algorithm is given that captures all these results, together with a sufficiency condition for its success. Time permitting, the application of HappyMap to computer vision and large language models will also be discussed.

9:30 **Shape analysis of functional data**
Karthik Bharath

Abstract: Qualitative descriptions of the shape of a function (e.g., convex, monotone, bimodal) and their use in developing and evaluating methodological tools abound in functional data analysis. The shape of a function is inextricably tied to its amplitude. A common approach to access shape information in a functional dataset is through a registration or alignment procedure to decouple amplitude from phase variations, which inevitably affects any downstream analysis. A hitherto unexplored alternative is to work directly with the shape space defined as the quotient of the function space under a group of shape-preserving transformations that treats phase as nuisance. I will discuss a stratified geometry for such a shape space that has regions of non-positive and (positive) unbounded curvature and discuss its statistical implications when analysing functional data.

10:00 **Selective inference with randomized Group LASSO estimators for general models**
Snigdha Panigrahi

Abstract: Our work is motivated by the need for inference after regularized estimation with high dimensional datasets that contain grouped covariates. As an example, consider applying a logistic Group LASSO to a dataset with a binary outcome and categorical predictors. How do we conduct selective inference in the estimated sparse model? This problem is challenging due to two reasons: (1) existing approaches for a polyhedral selection method do not apply to the Group LASSO because there is no easy description of the selection event; (2) our data is no longer normal. To solve this problem, we provide an asymptotic selective likelihood that uses extra randomization to obtain an easy to describe selection event. Our new approach provides selective inference using randomized Group LASSO estimators in likelihood models including generalized linear models, and in other general forms of estimation, such as quasi-likelihood estimation to include overdispersion, for example.

10:30 **Nonparametric density estimation from streaming data**
Aurore Delaigle

Abstract: We consider nonparametric density estimation from streaming data such as observations collected from sensor networks. Those data are characterized by their continuous collection over time in a high-velocity and often nonstationary environment, requiring near-real-time low-storage processing methods. We study the properties of an iterative estimator, which does not require storing data for long periods of time nor accessing them repeatedly. Then we suggest a procedure for implementing it in practice.

Causal inference in medical and public health studies

Organizer: Li Hsu
Chair: Yu Shen
Room: Sala Polivalente 1.1

9:00 **Estimation of the complier causal hazard ratio under dependent censoring**
Gilles Crommen, Jad Beyhum, Ingrid Van Keilegom

Abstract: In this work, we are interested in studying the causal effect of an endogenous binary treatment on a dependently censored duration outcome. By dependent censoring, it is meant that the duration time (T) and censoring time (C) are not statistically independent of each other, even after conditioning on the measured covariates. The endogeneity issue is handled by making use of a binary instrumental variable for the treatment. To deal with the dependent censoring problem, it is assumed that on the stratum of compliers: (i) T follows a semiparametric proportional hazards model; (ii) C follows a fully parametric model; and (iii) the relation between T and C is modeled by a parametric copula, such that the association parameter can be left unspecified. In this framework, the treatment effect of interest is the complier causal hazard ratio (CCHR). We devise an estimation procedure that is based on a weighted maximum likelihood approach, where the weights are the probability of an observation being a complier. The weights are estimated nonparametrically in a first stage, followed by the estimation of the CCHR. Conditions under which the model is identifiable are given, a two-step estimation procedure is proposed and some important asymptotic properties are established. Simulations are used to assess the validity and finite-sample performance of the estimation procedure. Finally, we apply the approach to estimate the causal effect of periodic screening examinations on time until death from breast cancer.

9:30 **A Multi-State Modeling for the Cost-Effectiveness Analysis in Disease Prevention**
Li Hsu

Abstract: Screening intervention has been widely recognized as an effective strategy for preventing chronic disease. Despite its effectiveness, determining when to start screening is complicated, because starting too early increases the number of screenings over lifetime and thus costs but starting too late may miss the cancer that could have been prevented. Therefore, to make an informed recommendation on the age to start screening, it is necessary to conduct cost-effectiveness analysis to assess the gain in life years relative to the cost of screenings. As more large-scale observational studies become accessible, there is growing interest in evaluating cost-effectiveness based on empirical evidence. In this talk, I will present a unified measure for evaluating cost-effectiveness and a causal analysis for the continuous intervention of screening initiation age, under the multi-state modeling with semi-competing risks. I will also present simulation results to show that the proposed estimators perform well in realistic scenarios. We perform a cost-effectiveness analysis of the colorectal cancer screening, utilizing data from the large-scale Women's Health Initiative. Our analysis reveals that initiating screening at age 50 years yields the highest quality-adjusted life years with an acceptable incremental cost-effectiveness ratio, providing real-world evidence in support of the US Preventive Services Task Force recommendation.

10:00 **Using Joint Models for Longitudinal and Time-to-Event Data to Investigate the Causal Effect of Salvage Therapy after Prostatectomy**
Jeremy Taylor, Dimitris Rizopoulos

Abstract: Prostate cancer patients who undergo prostatectomy are closely monitored for recurrence and metastasis using routine prostate-specific antigen (PSA) measurements. When PSA levels rise, salvage therapies are recommended in order to decrease the risk of metastasis. However, due to the side effects of these therapies and to avoid over-treatment, it is important to understand for which patients and when to initiate these salvage therapies. In this work, we use the University of Michigan Prostatectomy Registry Data to tackle this question. Due to the observational nature of this data, we face the challenge that PSA is simultaneously a time-varying confounder and an intermediate variable for salvage therapy. We define different causal salvage therapy effects defined conditionally on different specifications of the longitudinal PSA history. Specifically, for each patient, at each time they could receive salvage therapy, we define the causal effect for that patient as the difference in the probability of developing metastases within 2 years if they were to receive salvage therapy compared to not receiving salvage therapy. These effects are averaged over appropriate subsets of the patients to a marginal causal effect of salvage therapy. We then illustrate how these effects can be estimated using a Bayesian approach within the framework of joint models for longitudinal and time-to-event data.

10:30 **Causal and Statistical Uncertainty for Individual-Level Causal Inference**
Uri Shalit

Abstract: In recent years there has been growing interest in estimating individual-level causal effects, with applications in personalized medicine, economics, education and more. The quantity of interest for this task is known as the Conditional Average Treatment Effect (CATE), where the conditioning is on a high-dimensional set of covariates. The intended use of many of these methods is to inform human decision-makers about the probable outcomes of possible actions, for example, clinicians choosing among different medications for a patient. For such high-stakes decisions, it is crucial to responsibly convey a measure of uncertainty about its output, in order to enable informed decision making on the side of the human and to avoid catastrophic errors. We will present methods for estimating different sources of uncertainty in CATE models, focusing on the distinction between statistical uncertainty and causal uncertainty, and show how these measures of uncertainty can be used to responsibly decide when to defer decisions to experts and avoid unwarranted errors.

Statistics for AI
Organizer: Yongdai Kim
Chair: Yongdai Kim
Room: Sala Polivalente 1.2

9:00 **Minimax optimal density estimation using a shallow generative model**
Chae Minwoo, Hyeok Kyu Kwon

Abstract: A deep generative model yields an implicit estimator for the unknown distribution or density function of the observation. This paper investigates some statistical properties of the implicit density estimator pursued by VAE-type methods from a nonparametric density estimation framework. More specifically, we obtain convergence rates of the VAE-type density estimator under the assumption that the underlying true density function belongs to a locally Holder class. Remarkably, a near minimax optimal rate with respect to the Hellinger metric can be achieved by the simplest network architecture, a shallow generative model with a one-dimensional latent variable.

9:30 **A statistical analysis of an image classification problem**
Sophie Langer, *Juntong Chen*, Johannes Schmidt-Hieber

Abstract: The availability of massive image databases resulted in the development of scalable machine learning methods such as convolutional neural network (CNNs) filtering and processing these data. While the very recent theoretical work on CNNs focuses on standard nonparametric denoising problems, the variability in image classification datasets does, however, not originate from additive noise but from variation of the shape and other characteristics of the same object across different images. To address this problem, we consider a supervised classification problem for object detection on grayscale images. While from the function estimation point of view, every pixel is a variable and large images lead to high-dimensional function recovery tasks suffering from the curse of dimensionality, increasing the number of pixels in our image deformation model enhances the image resolution and makes the object classification problem easier. We propose and theoretically analyze two different procedures. The first method estimates the image deformation by support alignment. Under a minimal separation condition, it is shown that perfect classification is possible. The second method fits a CNN to the data. We derive a rate for the misclassification error depending on the sample size and the number of pixels.

10:00 **Statistical Analysis on In-Context Learning**
Masaaki Imaizumi

Abstract: Deep learning and artificial intelligence technologies, one of the modern data science technologies, have made great progress, and their mathematical understanding is required to efficiently control and develop these technologies. In this talk, we present a statistical analysis of a scheme called in-context learning, which explains foundation models for artificial intelligence such as ChatGPT. We argue that in-context learning can achieve efficient learning under certain conditions, owing to the property of the transformer, which can handle the various property of empirical distributions.

10:30 **Optimal high-dimensional nonparametric regression with variational neural networks**
Ilsang Ohn

Abstract: In this work, we study high-dimensional nonparametric regression problems, where the number of predictors is allowed to diverge as the sample size grows. We introduce a new class of sparse regression models that assume the true regression function consists of several component functions, each of which lies on a subspace spanned by a small number of predictors. This function class includes a standard sparse regression model, a sparse additive model and a multi-index model as special cases. We propose to use variational neural networks with continuous spike-and-slab priors, which offer efficient computation. We prove that a variational deep neural network attains a minimax optimal contraction rate under this model. Notably, the contraction rate is simultaneously adaptive to both the unknown smoothness and sparsity of the regression function.

Statistics for dependent data
Organizers: Jens-Peter Kreiss and Efstathios Paparoditis
Chair: Jens-Peter Kreiss and Efstathios Paparoditis
Room: Sala Polivalente 1.3

9:00 **A log-linear model for non-stationary time series of counts**
Michael H. Neumann, Anne Leucht

Abstract: We propose a new model for nonstationary integer-valued time series which is particularly suitable for data with a strong trend. In contrast to popular Poisson-INGARCH models, but in line with classical GARCH models, we propose to pick the conditional distributions from nearly scale invariant families where the mean absolute value and the standard deviation are of the same order of magnitude. As an important prerequisite for applications in statistics, we prove absolute regularity of the count process with exponentially decaying coefficients. This talk is based on joint work with Anne Leucht (Universität Bamberg).

9:30 **Autoregressive Network: Sparsity and Degree Heterogeneity**
Yutong Wang

Abstract: This paper proposes a new dynamic network model with sparsity and degree heterogeneity. Sparsity means the expected edge density tending to 0, while degree heterogeneity captures the node-level difference in forming and dissolving an edge. We develop a new concentration inequality for the dependent sequence while the alpha-mixing coefficient goes to 0 as the number of nodes p goes to infinity. We proposed a two-step estimation strategy, where the first step is the Maximum Likelihood Estimation by ignoring the degree heterogeneity structure, and the second step is an M-estimation based on the maximum likelihood estimator in the first step. To evaluate the estimator's performance, we give theoretical results regarding the upper bound for the estimation error for the proposed estimator. The theory of the model is supported by extensive simulation.

10:00 **Learning Graphical Models for nonstationary multivariate time series**
Suhasini Subba Rao, Jonas Krampe

Abstract: In a recent paper we introduced NonStGM, a general nonparametric graphical modeling framework for studying dynamic associations among the components of a nonstationary multivariate time series. It builds on the framework of Gaussian Graphical Models (GGM) and stationary time series Graphical models (StGM), Analogous to StGM, the proposed framework captures conditional noncorrelations in the form of an undirected graph. In addition, to describe the more nuanced nonstationary relationships among the components of the time series, the new notion of conditional nonstationarity/stationarity is incorporated within the graph architecture. In the current talk we propose a method for estimating the graph architecture from data. We show that the notions described above are succinctly encoded in the precision matrix of the Discrete Fourier Transform of the time series. Therefore, we propose a method for consistently estimating this high dimensional precision matrix. We estimate the corresponding graph by testing for conditional correlation and conditional stationarity using the entries of the estimated precision matrix. We present the sampling properties of the proposed estimators and illustrate our method with real data.

10:30 **Asymptotic Theory for Constant Step Size Stochastic Gradient Descent**
Jiaqi Li, Wei Biao Wu, Zhipeng Lou, *Stefan Richter*

Abstract: In this talk, we present a novel approach to understanding the behavior of Stochastic Gradient Descent (SGD) with constant step size by interpreting its evolution as a Markov chain. Unlike previous studies that rely on the Wasserstein distance, our approach leverages the functional dependence measure and explore the Geometric-Moment Contraction (GMC) property to capture the general asymptotic behavior of SGD in a more refined way. In particular, our approach allow SGD iterates to be non-stationary but asymptotically stationary over time, providing quenched versions of the central limit theorem and invariance principle valid for averaged SGD with any given starting point. We subsequently define a Richardson-Romberg extrapolation with an improved bias representation to bring the estimates closer to the global optimum. We establish the existence of a stationary solution for the derivative SGD process under mild conditions, enhancing our understanding of the entire SGD procedure. Lastly, we propose an efficient online method for estimating the long-run variance of SGD solutions.

Object Oriented Data Analysis: Trees and Graphs

Organizer: Steve Marron
Chair: Stephan Huckemann
Room: Sala Polivalente 1.4

9:00 **Barycentric Subspace Analysis for Sets of Unlabelled Graphs**
Anna Calissano, Elodie Maignant, Xavier Pennec

Abstract: Unlabelled graphs represent those scenarios in which there is no clear correspondence between the nodes across the graphs. Such graphs arises in different applications and unlabelled graphs encode the most general case when working with sets of graphs. In this talk, we introduce Barycentric Subspace Analysis (BSA) for unlabelled graphs, as a powerful and interpretable technique for dimensionality reduction. Identifying each graph by the set of its eigenvalues, the graph spectrum space is defined as a novel and computationally efficient quotient manifold of isospectral graphs. In such a manifold, the notion of BSA is extended. We showcase how BSA can be used as a powerful technique for dimensionality reduction of complex data. In details, BSA searches for a subspace of a lower dimension, minimising the projection of data points onto such subspace. As the subspace is identified by a set of reference points, the interpretation is easier than with other dimensionality reduction techniques. BSA is performed and compared with clustering and PCA on a simulated dataset and a real-world dataset of airline company networks.

9:30 **Sticky Flavors**
Stephan F. Huckemann, Lars Lammers, Do Tran Van

Abstract: Sample spaces modeling phylogenetic trees (such as BHV space or wald space) intrinsically carry a stratified structure featuring infinite negative curvatures: too many less resolved tree topologies meet at higher resolved tree topologies, preventing a manifold structure there. In result probability distributions of phylogenetic trees may feature stickiness of their Fréchet means: after a finite random sample size, the asymptotic distribution of sample means collapses. In the worst case, they collapse toward a single point, posing serious complications for asymptotic nonparametric statistics. In this talk we take a closer look at this phenomenon, exploring the various flavors of stickiness and their interdependence. For instance sample stickiness (above), perturbation and topological (e.g. total variation or Wasserstein) stickiness, as well as modulation and directional stickiness. The latter promise to be of value where current asymptotic nonparametric statistical method struggle or even fail.

10:00 **Brownian motion, bridges and Bayesian inference in BHV tree space**
Tom Nye, *William Woodman*

Abstract: Samples of phylogenetic trees arise in many different contexts, and it is important to be able to perform statistics on such data sets, for example calculating means and performing hypothesis tests. Billera-Homes-Vogtmann (BHV) tree space encapsulates the set of all possible trees on a fixed set of leaves, and a number of different statistical methods have been developed in BHV tree space, mostly based on least-squares approaches. We introduce probability distributions on BHV tree space which are parameterized by a location parameter and dispersion parameter. These are obtained by executing Brownian motion from a fixed starting point in tree space (determined by the location parameter) for a certain time duration (corresponding to the dispersion parameter). Approximate samples can be drawn from such distributions by simulating random walks. Fitting these distributions to samples of points in tree space is challenging, as the likelihood cannot be evaluated directly due to the complex geometry of BHV tree space. We model data sets of phylogenetic trees as independent draws from a random walk distribution. Brownian bridges are random walk trajectories conditioned on both endpoints. We augment the parameter space with Brownian bridges between the location parameter point and the data points. We then describe an approach to simulate from the Bayesian posterior of all the parameters using MCMC. Simulating the bridges relies on an efficient algorithm for proposing random walk trajectories between fixed start and end points in tree space. The algorithms yield an efficient Bayesian approach to fitting Brownian motion transition distributions to samples of trees in BHV tree space. Finally, marginal likelihoods can be calculated in order to compare different source parameters for data sets.

10:30 **Estimating a mean tree for phylogenetic trees with missing taxa**
Maryam Garba, Tom Nye

Abstract: Most methods for analysing samples of phylogenetic trees assume the trees all have the same set of taxa at the leaves, but this is not the case for many genetic data sets. The recently-developed wald space of phylogenetic trees is obtained via an embedding of trees into the space of positive definite matrices. While the wald space similarly assumes all trees have the same set of leaves, the embedding enables analysis of trees with missing taxa via the ambient affine invariant metric. We describe methods for computing an extrinsic Fréchet mean for a sample of trees with different taxa, which is then projected to obtain a mean tree. We study properties of this estimator via a simulation study.

Analysis of curves
Organizer: Aurore Delaigle
Chair: Aurore Delaigle
Room: Sala Polivalente 1.5

9:00 **Multivariate higher-order kernels**
José E. Chacón, Tarn Duong

Abstract: Higher-order kernels offer improved statistical properties in nonparametric kernel smoothing, such as reduced bias, compared to second-order kernels. In the multivariate case, most of the existing methodologies for defining higher-order kernels focus on product kernels, despite their well-known sub-optimal performance as compared to spherically symmetric kernels. Surely one of the reasons for this admittedly limited approach is the lack of a systematic framework to treat higher-order derivatives of multivariate functions. By employing a novel vectorized differential analysis framework, here the well-established recursive rules to generate higher-order univariate kernels are extended to multivariate spherically symmetric kernels. Furthermore, closed-form non-recursive expressions for multivariate Gaussian-based higher-order kernels are also exhibited.

9:30 **Robust estimation under small measurement errors**
Michael Stewart, Alan Welsh

Abstract: Consider estimating the scale of the random effect distribution in a balanced, one-way random effects model. Classically, the variance component is estimated using the cluster means. If we wish more robust methods, how do we proceed? The cluster means may be viewed as "true" random effects contaminated with measurement error. If we consider an asymptotic scheme whereby the cluster sizes increase with the number of clusters, the dispersion of the measurement errors is estimable and decreasing. We thus return to the "small measurement error" scenario of Stefanski (Biometrika, 1985) where bias-reduction is the main focus. We compare various bias-reduction strategies which involve estimating a score function as an intermediate step and demonstrate how adding some extra smoothing can lead to better results once appropriate bias-correction is implemented.

10:00 **Partially observed functional data over non-Euclidean domains**
Alessandro Palummo, Marco Stefanucci, Eleonora Arnone, *Laura M. Sangalli*

Abstract: I will discuss an innovative class of Physics-Informed statistical learning methods for the analysis of functional data observed over multidimensional non-Euclidean domains, such as two-dimensional manifolds and non-convex volumes. These methods, including Functional Principal Component Analysis, can handle sparse and partially observed functional data, as well as massive data. I will illustrate the methods with challenging applications to environmental problems.

10:30 **Kernel estimation for continuous-time semi-Markov processes**
Vlad Stefan Barbu

Abstract: Our presentation is dedicated to kernel estimation some characteristics of a continuous-time semi-Markov process, like sojourn time distributions in a state, semi-Markov kernel, Markov renewal and semi-Markov transition functions. We construct nonparametric kernel estimators and we establish asymptotic properties of these estimators, when the sample size becomes large. The qualities of the estimators are illustrated by a numerical example. This is a joint work with Chafiâa AYHAR (University Center of El Bayadh Nour El Bachir, Algeria; ayharchafiaa@yahoo.com), Fatiha MOKHTARI (LSMSA, University of Saida–Doctor Moulay Taher, Algeria; fatiha.mokhtari@univ-saida.dz), Saâdia RAHMANI (LSMSA, University of Saida–Doctor Moulay Taher, Algeria; saadia.rahmani@univ-saida.dz). [1] C. Ayhar, V. S. Barbu, F. Mokhtari, S. Rahmani, On the asymptotic properties of some kernel estimators for continuous semi-Markov processes, Journal of Nonparametric Statistics, 34(2), 299-318, 2022 [2] F. Mokhtari, C. Ayhar, V. S. Barbu, S. Rahmani, Kernel estimators of Markov renewal and semi-Markov transition functions of semi-Markov systems, 2024, under revision

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk 2
Chair: Malka Gorfine
Room: Grande Auditorio

11:30 **Deep Learning for Censored Survival Data**
Jane-Ling Wang

Abstract: Unlike standard tasks, survival analysis requires modeling incomplete data, such as right-censored data, which must be treated with care. While deep neural networks excel in traditional supervised learning, it remains unclear how to best utilize these models in survival analysis. A key question asks which data-generating assumptions of traditional survival models should be retained and which should be made more flexible via the function-approximating capabilities of neural networks. In addition, most of these methods are difficult to interpret and mathematical understanding of them is lacking. In this talk, we explore these issues from two directions. First, we study the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. Next, we introduce a Deep Extended Hazard (DeepEH) model to provide a flexible and general framework for deep survival analysis. The extended hazard model includes the conventional Cox proportional hazards and accelerated failure time models as special cases, so DeepEH subsumes the popular Deep Cox proportional hazard (DeepSurv) and Deep Accelerated Failure Time (DeepAFT) models. We provide theoretical support for the proposed models, which underscores the attractive feature that deep learning is able to detect low-dimensional structure of data in high-dimensional space. Numerical experiments further provide evidence that the proposed methods outperform existing statistical and deep learning approaches to survival analysis. Time permitting, we will explore how to perform hypothesis testing for survival data. This is joint work with Qixian Zhong (Xiamen University) and Jonas Mueller (Clean Lab).

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 6

Current topics in biostatistics - nonparametric approaches
Organizer: Somnath Datta
Chair: Michael Daniels
Room: Grande Auditorio

13:30 **Analysis of spatially clustered survival data with unobserved covariates using SBART**
Debajyoti Sinha, Durbadal Ghosh, Antonio Linero, George Rust

Abstract: Popular parametric and semi-parametric regression methods for clustered survival data are inappropriate and inadequate when the appropriate functional forms of the covariates and their interactions are unknown, and random cluster effects as well as some unknown cluster-level covariates are spatially correlated. We present a general nonparametric method for such data under a paradigm of Bayesian ensemble learning called Soft Bayesian Additive Regression Trees (SBART in short). Our additional methodological and computational challenges challenges include a large number of clusters, variable cluster sizes, and data information for proper statistical augmentation of the unobserved covariate being sourced from a data registry different from the survival study. We illustrate the practical implementation of our method and its advantages over existing methods via assessing the impacts of intervention in some cluster/county level and patient-level covariates to eliminate existing racial disparity in breast cancer survival in different Florida counties (clusters), where the clustered survival data with patient-level covariates come from Florida Cancer Registry (FCR), and the data information for one unobserved county-level covariate come from the Behavioral Risk Factor Surveillance Survey (BRFSS). We also compare our method with various existing semiparametric analysis methods to demonstrate our advantages via simulation studies.

14:00 **A Bayesian nonparametric approach for nonignorable missingness in EHR data**
Michael Daniels, David Lindberg, Sebastien Haneuse

Abstract: We propose an approach for missingness in EHRs using Bayesian nonparametric (BNP) models. We show how to introduce sensitivity parameters corresponding to nonignorable missingness in the outcome and confounders by extracting unidentified distributions from the BNP model and reconstructing the distribution of interest. We also flexibly include auxiliary covariates to move closer to MAR. We use G-computation based on the reconstructed distribution to compute causal estimands of interest. We use our approach to assess the comparative effectiveness of two bariatic surgeries on BMI 18 months after surgery.

14:30 **Bayesian Nonparametric Modeling of Restricted Mean Survival Time: Subject Specific Inference and Average Treatment Effect**
Sanjib Basu, Ruizhe Chen

Abstract: Restricted mean survival time (RMST) is increasingly being used in planning and analyzing time-to-event outcome in clinical, medical and health sciences. The concept of RMST is model independent and its popularity stems from its advantages as a summary in survival analysis over other conventional measures such as the hazard ratio. We develop a Bayesian nonparametric model for RMST and develop subject-level RMST inference as well as group-level causal inference of average treatment effect. We evaluate performance of the proposed Bayesian nonparametric approach and compare with non-Bayesian methods. We present an application of our proposed approach to analyze data from a metastatic colorectal cancer trial.

15:00 **Error Controlled Feature Selection for Ultrahigh Dimensional and Highly Correlated Feature Space Using Deep Learning**
Taps Maiti

Abstract: Deep learning has been at the center of analytics in recent years due to its impressive empirical success in analyzing complex data objects. Despite this success, most existing tools behave like black-box machines, thus the increasing interest in interpretable, reliable, and robust deep learning models applicable to a broad class of applications. Feature-selected deep learning has emerged as a promising tool in this realm. However, the recent developments do not accommodate ultra-high dimensional and highly correlated features or high noise levels. In this article, we propose a novel screening and cleaning method with the aid of deep learning for a data-adaptive multi-resolutional discovery of highly correlated predictors with a controlled error rate. Extensive empirical evaluations over a wide range of simulated scenarios and several real datasets demonstrate the effectiveness of the proposed method in achieving high power while keeping the false discovery rate at a minimum.

Regularized nonparametric regression for spatial and functional data
Organizer: Laura M. Sangalli
Chair: Eleonora Arnone
Room: Pequeno Auditorio

13:30 **Sparsistency of estimators in semiparametric mixture of regression models**
Abbas Khalili

Abstract: Semiparametric finite mixture of varying coefficient regressions (FM-VCR) provide a rich class of statistical models for capturing unobserved heterogeneity in the data while accounting for heterogeneous varying covariates' effects on a response variable. Although complex, this situation frequently occurs in real data applications as illustrated through the analysis of a genetic dataset in our demonstration. Oftentimes, the number of covariates with (nonparametric) varying coefficients also presents a challenge. In this talk, we will introduce regularized local likelihood methods for simultaneous parameter estimation and variable selection in sparse FM-VCR models. Furthermore, we discuss sparsistency property exhibited by these estimators. To showcase the efficacy of these techniques, a simulation study will be presented. Lastly, we will explore the application of sparse FM-VCR models in the analysis of osteocalcin (OCN) data for identifying genetic factors with age-dependent effects on OCN in a nonparametric fashion.

14:00 **A flexible framework for spatial quantile regression via PDE regularization**
Cristian Castiglione

Abstract: Differential penalization driven by partial differential equations (PDE) represents an innovative, powerful tool for incorporating physical information and structured smoothing into complex nonparametric spatial regression problems. It is suited for the analysis of spatial and functional data observed over complicated domains and characterized by non-trivial dependence patterns. To flexibly model non-Gaussian, heteroscedastic and possibly skew-distributed environmental data, we leverage such an approach by proposing a flexible quantile regression model with PDE regularization. We present an efficient estimation algorithm and a novel cross-validation criterion tailored for smoothing parameter selection. To ensure numerical tractability, we employ a suitable discretization method based on finite element techniques. Expanding beyond the foundational framework, we extend our methodology to embrace spatio-temporal and spatial functional data, possibly contaminated by complex missing value patterns. Additionally, we propose an innovative method for jointly estimating multiple quantiles, enhancing the model's applicability for a nonparametric reconstruction of the whole conditional distribution of the spatial phenomena under study. The proposed spatial quantile regression framework is completely implemented and integrated within the computational infrastructure provided by the R package fdaPDE, which provides an effective, user-friendly interface for the analysis of complex spatial data.

14:30 **A regularized compositional functional concurrent regression model to investigate the dynamic relationship between causes of death and human longevity**
Emanuele Giovanni Depaoli, *Marco Stefanucci*, Stefano Mazzuco

Abstract: In this study, we introduce a novel functional concurrent regression model where independent variables are functional compositions, a data structure which is attracting increasing interest in the statistical modeling literature. This framework allows for the exploration of temporal relationships between life expectancy at birth and compositions derived from cause-specific mortality rates across four distinct age classes in a sample of countries. We propose a penalized approach for estimating regression coefficients and selecting relevant variables. Subsequently, we devise an efficient computational strategy based on an augmented Lagrangian algorithm to solve the resulting optimization problem. Through a simulation study, we demonstrate the model's effectiveness in predicting the response function and estimating unknown functional coefficients. Analysis of real data reaffirms the significant role of neoplasms and cardiovascular diseases in determining life expectancy, as identified in prior studies, while also unveiling additional contributions not previously observed.

15:00 **Function Estimation on Complex 3D Surfaces**
Michelle Carey, Thiago Da Silva Cardoso

Abstract: Medical imaging techniques like Functional Magnetic Resonance Imaging (fMRI) are indispensable in modern medicine for their non-invasive ability to study the human brain. Recent advancements in medical software can now translate fMRI scans into detailed three-dimensional digital brain meshes, incorporating anatomical measurements. However, analyzing such data poses challenges due to the brain's intricate non-Euclidean geometry and the presence of noise from the imaging process. This study introduces an innovative method to estimate a smooth underlying signal from noisy data associated with 3D brain surfaces. The approach integrates concepts from deep learning and Functional Data Analysis (FDA) to develop physics-informed neural networks. These networks are trained to effectively smooth data over complex surfaces. Simulations reveal that our model outperforms existing methods by up to 50%. Finally, we apply this method to investigate the resting-state brain activity levels of a healthy individual.

Recent advances in spatiotemporal data

Organizer: George Michailidis
Chair: George Michailidis
Room: Sala Polivalente 1.1

13:30 **Semi-Parametric Inference for Doubly Stochastic Spatial Point Processes: An Approximate Penalized Poisson Likelihood Approach**
Ali Shojaie, Si Cheng, Jon Wakefield

Abstract: Doubly-stochastic point processes model the occurrence of events over a spatial domain as an inhomogeneous Poisson process conditioned on the realization of a random intensity function. They are flexible tools for capturing spatial heterogeneity and dependence. However, implementations of doubly-stochastic spatial models are computationally demanding, often have limited theoretical guarantee, and/or rely on restrictive assumptions. We propose a penalized regression method for estimating covariate effects in doubly-stochastic point processes that is computationally efficient and does not require a parametric form or stationarity of the underlying intensity. Our approach is based on an approximate (discrete and deterministic) formulation of the true (continuous and stochastic) intensity function. We show that consistency and asymptotic normality of the covariate effect estimates can be achieved despite the model misspecification, and develop a covariance estimator that leads to a conservative statistical inference procedure. A simulation study shows the validity of our approach under less restrictive assumptions on the data generating mechanism, and an application to Seattle crime data demonstrates better prediction accuracy compared with existing alternatives.

14:00 **Likelihood Free Learning of Saptiotemporal Hawkes Processes**
Moulinath Banerjee

Abstract: Likelihood Free Learning of Spatiotemporal Hawkes Processes Abstract: Hawkes Processes are quite popular for analyzing spatiotemporal data with triggering effects and have been used as a tool for algorithmic threat detection. However, in real applications, complete data on sample paths are usually unavailable (e.g. unreported crime), whilst (estimates of) missing rates may be known. As the intensity function of a Hawkes process depends on past events, this makes the use of likelihood based methods like EM essentially infeasible. On the other hand, MDE (minimum distance estimates) based on Wasserstein distances are readily computable using GAN training, as samples from a Hawkes process with a fixed set of parameters can be readily generated. We illustrate the use of such MDE estimates to learn the parameters of Hawkes processes and present applications to predictive policing. We also investigate the theoretical properties of the estimators by invoking recent work on entropy regularized optimal transport theory. This is joint work with Pramit Das, Yuekai Sun and Yue Yu.

14:30 **Impulse Response Estimation in Large-scale Time Series**
Sumanta Basu

Abstract: Impulse Response Function (IRF) estimation is a canonical problem in multivariate time series. Impulse responses capture how shocks applied to one component of a multivariate dynamical system propagate to its other components over time. In the high-dimensional regime, majority of existing works focus on a sparse vector autoregressive (VAR) model specification. In some applications, however, imposing sparsity constraints directly on the space of impulse responses can provide a more interpretable description of IRF. In this work, we adopt a sparse vector moving average (VMA) model specification to estimate impulse responses in high-dimensional time series. We propose an iterative algorithm for learning cumulative impulse response functions, and demonstrate how this can be used to build graphical models for large-scale dynamical system. We provide asymptotic analysis of our proposed method, and illustrate its advantages over competing alternatives using simulated and a real data set from financial economics.

15:00 **Clustering of spatiotemporal processes using spectral analysis and applications**
Soudeep Deb, *Sayar Karmakar*

Abstract: In this talk I present a new clustering algorithm for spatio-temporal data. The proposed method leverages a weighted combination of a spatial haversine distance matrix and a spectral- density based temporal distance matrix between the locations. Concepts of partition around medoids algorithm and the gap statistic are utilized to develop the algorithm and to determine the optimal number of clusters. Such a non-parametric algorithm is novel as it incorporates both spatial and temporal distances of the units and it can work for time-series of possibly different lengths. Theoretical guarantee of consistency of the proposed method is provided. An elaborate simulation study is also given to demonstrate the efficacy of the algorithm. As an interesting real life application, the proposed algorithm is implemented to analyze the spatio-temporal dynamics of the time series of coronavirus (COVID-19) incidence rates observed at county-level in the United States of America. The results are demonstrated on datasets of different sizes: the entire country, the Midwest region and the state of California. Special emphasis is given on the last two cases to display how the clustering results offer interesting insights into the epidemic progression in these areas. Particularly, it sheds light on whether state-mandated restrictions impacted the entire state similarly or if there are interesting local behaviors in terms of the COVID-19 spread. We conclude with some related current works we are doing on spatiotemporal data.

Bayesian nonparametrics for complex data

Organizer: Ramses Mena
Chair: Ramses Mena
Room: Sala Polivalente 1.2

13:30 **Finite population inference via martingales with a view towards quick counts**
Carlos E. Rodríguez

Abstract: Statistical inference in a finite population setting is of interest in many areas where one seeks to quantify the uncertainty about the unobserved members of the population based on those that are observed. Though the literature comprises some frequentist and Bayesian solutions in such a framework, these are mostly tailored for specific applications. Here, the novel approach of martingale predictive inference is explored and used to quantify uncertainty for a population statistic of interest in various models for finite populations. The central idea resides in a martingale property of estimators. Motivated by the Quick Count organized by the National Electoral Institute of Mexico, {two real-world datasets are used to test our proposal}: the 2021 Mexican referendum to determine whether former presidents can be prosecuted for corruption and voting data from the 2017 governor election in the State of Mexico. The proposal is also confronted and discussed against other methods available in the literature. Supporting information for this article, including the code and datasets used for the analysis, is available online.

14:00 **Efficient estimation of the Posterior Similarity Matrix for Bayesian Nonparametric clustering**
Johan van der Molen Moris

Abstract: Mixture models, and in particular Dirichlet Process mixture models, are widely used in Bayesian model-based clustering. The estimation of the corresponding posterior distribution is typically done using Markov Chain Monte Carlo (MCMC) methods. In this context, the Posterior Similarity Matrix (PSM) is crucial for obtaining a point estimate of the clustering structure. This matrix summarises the MCMC samples of the cluster labels, providing a comprehensive overview of the data's structure. Despite the importance of the PSM, the current MCMC methods used to estimate it present several limitations. They can be slow to converge, especially in high-dimensional spaces, and need careful initialization. Moreover, checking the convergence of MCMC chains is not trivial and can lead to wrong inference if not done carefully. MCMC chains typically get stuck at a single partition of the data, or they visit a very small number of them. This results in a restricted version of the posterior distribution that can affect negatively the PSM estimation. In this work, a novel approach is proposed to solve these problems by providing a faster and more efficient estimation of the PSM without the use of MCMC. This method is based on an algorithm that approximates the posterior similarity matrix entries directly, leveraging an analytical expression, in the case of a conjugate Dirichlet Process mixture model. This not only reduces the computational cost but also improves the accuracy of the estimation of the matrix, thus providing a more precise representation of the posterior distribution. In this presentation, I will go into the details of the proposed method, and show results on simulated and real data, illustrating its advantages over traditional MCMC methods, as well as some of its own challenges and possible extensions, particularly in the context of big data.

14:30 **Clustering constrained on linear networks**
Asael Fabian Martinez Martinez

Abstract: Motivated by the analysis of georeferenced crime data over the streets in Mexico City, a simple method for clustering observations living on a linear network is presented. This method is based on the Dirichlet process and penalizes the grouping of points according to their distance. The performance of the proposal is illustrated by using simulated and real data.

15:00 **Conditional partial exchangeability: a probabilistic framework for longitudinal and multi-view clustering**
Beatrice Franzolini

Abstract: Clustering longitudinal data requires adjusting the number of clusters, their frequencies, and shapes over time to accurately capture heterogeneity and temporal dynamics. However, many existing dynamic clustering techniques overlook within-subject dependence, thereby neglecting the identities of subjects over time. This issue is also encountered in stochastic block models for longitudinal and multiplex network data. To overcome these limitations, we propose a broad class of Bayesian mixture models capable of generating dependent random partitions, where the dependency is introduced at the subject level. At the core of our proposal is conditional partial exchangeability. This novel probabilistic paradigm ensures analytical and computational tractability while defining a flexible law governing dependent random partitions of the same objects across time, space, or domains.

Nonparametric statistics: methods and applications

Organizer: Sonali Das
Chair: Sonali Das
Room: Sala Polivalente 1.3

13:30 **Co-variance Operator of Banach Valued Random Elements: U-Statistic Approach**
Subhra Sankar Dhar

Abstract: In this talk, we propose a co-variance operator for Banach valued random elements using the concept of U-statistic. We then study the asymptotic distribution of the proposed co-variance operator along with related large sample properties. Moreover, specifically for Hilbert space valued random elements, the asymptotic distribution of the proposed estimator is derived even for dependent data under some mixing conditions. This is a joint work with Suprio Bhar (IIT Kanpur, India).

14:00 **On A Goodness-of-fit Test for Elliptically Symmetric Distributions based on Scale-Scale Plots**
Biman Chakraborty, Pritha Guha

Abstract: In this talk, we propose a test of goodness of fit for families of elliptically symmetric distributions based on a test statistic derived from scale–scale plots. These scale-scale plots can be viewed as a multivariate analog of quantile-quantile plots, which are constructed uisng the volume functionals of the central rank regions. The test is motivated through the multivariate normal distributions and extended to a test of elliptical symmetry. We derive the asymptotic properties of the test statistic, and perform detailed power studies for the test of goodness of fit, as well as the test for elliptical symmetry. We also compare the performance of the proposed test with some well-known alternatives.

14:30 **Are winters getting shorter?**
Anandamayee Majumdar, Sonali Das, Mehmet Balcilar, Levi Baguley, Siphumile Mangisa

Abstract: A recurrent concern from agricultural scientists and ecologists is that 'winters are becoming shorter and warmer' and is noticeably affecting both the physical and biological systems, and manifesting in different concerning ways that are impacting lives and livelihoods. The changing features of seasons have natural consequences not only on the overall ecosystem of living beings, but also on the way economic decisions are made and on their outcomes. This research aims to analyze long-term temperature data to assess if winters are indeed becoming shorter, and also warmer. This research explores both hemispheres and uses data in US and South Africa. We will use exploratory data analysis and nonparametric statistical methods applied to temperature data. This research hopes to contribute valuable insights to ecological and biological resource management.

15:00 **Nonparametric Quantile Causality Assessment of Uncertainty and Gold: Multivariate and Bootstrap Extensions**
Mehmet Balcilar, Rangan Gupta

Abstract: This study extends the nonparametric causality-in-quantile test approach to multivariate settings that uses bootstrapping. We apply the tests to study whether gold acts as a hedge against different forms of economic and financial uncertainty. Gold is commonly regarded as a "safe-haven" asset, implying that investors often turn to it in times of economic and geopolitical uncertainty. While gold is recognized as a safe-haven and considered a hedge against various financial asset risks, its effectiveness as a hedge for uncertainties appears ambiguous. Our study makes a significant contribution to the literature by providing robust results based on a diverse set of uncertainty measures.

Recent advances in depth and robust statistics

Organizer: Graciela Boente
Chair: Alicia Nieto-Reyes
Room: Sala Polivalente 1.4

13:30 **Robust Functional Regression with Discretely Sampled Predictors**
Ioannis Kalogridis

Abstract: The functional linear model is an important extension of the classical regression model allowing for scalar responses to be modeled as functions of stochastic processes. Yet, despite the usefulness and popularity of the functional linear model in recent years, most treatments, theoretical and practical alike, suffer either from (i) lack of resistance towards the many types of anomalies one may encounter with functional data or (ii) biases resulting from the use of discretely sampled functional data instead of completely observed data. To address these deficiencies, we introduce and study a class of robust functional regression estimators for partially observed functional data. The proposed broad class of estimators is based on thin-plate splines with a novel computationally efficient quadratic penalty, is easily implementable and enjoys good theoretical properties under weak assumptions. We show that, in the incomplete data setting, both the sample size and discretization error of the processes determine the asymptotic rate of convergence of functional regression estimators and the latter cannot be ignored. These theoretical properties remain valid even with multi-dimensional random fields acting as predictors and random smoothing parameters. The effectiveness of the proposed class of estimators in practice is demonstrated by means of a simulation study and a real-data example.

14:00 **Local depth functions and clustering**
Claudio Agostinelli, Giacomo Francisci, Anand Vidyashankar, Alicia Nieto-Reyes

Abstract: Statistical local depth functions are a generalization of statistical depth functions and they are used for describing the local geometric features and mode(s) in multivariate distributions. We illustrate some analytical and statistical properties of these functions. We show that, when the underlying probability distribution is absolutely continuous with density, an appropriate scaled version of local depth (referred to as tau-approximation) converges, uniformly and in L^d(R^p) to the density when the parameter controlling the system of neighborhoods converges to zero. We also establish that, as the sample size diverges to infinity the centered and scaled sample local depth converges in distribution to a centered Gaussian process uniformly in the space of bounded functions on a class of functions yielding local depths. We present, using the sample version of the tau-approximation and the gradient system analysis a new clustering algorithm. For this algorithm we discuss its consistency. We use applications to mode estimation and upper level set estimation to illustrate the proposed methods.

14:30 **Multivariate Singular Spectrum Analysis by Robust Diagonalwise Low-Rank Approximation**
Mia Hubert, Fabio Centofanti, Peter Rousseeuw

Abstract: Multivariate Singular Spectrum Analysis (MSSA) is a powerful and widely used nonparametric method for multivariate time series, which allows the analysis of complex temporal data from diverse fields such as finance, healthcare, ecology, and engineering. However, MSSA lacks robustness against outliers because it relies on the singular value decomposition, which is very sensitive to the presence of anomalous values. MSSA can then give biased results and lead to erroneous conclusions. We propose a new MSSA method, named RObust Diagonalwise Estimation of SSA (RODESSA), which is robust against the presence of cellwise and casewise outliers. In particular, the decomposition step of MSSA is replaced by a new robust low-rank approximation of the trajectory matrix that takes its special structure into account. A fast algorithm is constructed, and it is proved that each iteration step decreases the objective function. In order to visualize different types of outliers, a new graphical display is introduced, called an enhanced time series plot. An extensive Monte Carlo simulation study is performed to compare RODESSA with competing approaches in the literature. A real data example about temperature analysis in passenger railway vehicles demonstrates the practical utility of the proposed approach.

15:00 **On depth based two-sample tests: robustness in functional spaces**
Alicia Nieto-Reyes, Felix Gnettner, Claudia Kirch

Abstract: Statistical depth functions are used to order the elements of a space with respect to a distribution. They are applicable to spaces of any dimension, such as multivariate and functional spaces. Liu and Singh introduced in 1993 a depth-based multivariate two-sample test. This talk is mainly dedicated to improving the power of the associated test statistic, and making it symmetric in both samples. Additionally, we incorporate its applicability to functional data and provide some simulations where we analyze its robustness in comparison to other tests.

Cutting-edge machine learning for complex biomedical data
Organizer: Malka Gorfine
Chair: Malka Gorfine
Room: Sala Polivalente 1.5

13:30 **Deep Learning of Partially Linear Cox Models: Error Rate and Selection Consistency**
Yi Li

Abstract: Lung cancer is a leading cause of cancer mortality globally, highlighting the importance of understanding its mortality risks to design effective patient-centered therapies. The National Lung Screening Trial (NLST) employed computed tomography texture analysis, which provides objective measurements of texture patterns on CT scans, to quantify the mortality risks of lung cancer patients. Partially linear Cox models have gained popularity for survival analysis by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerging risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. We propose a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC), which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. We prove the convergence and asymptotic properties of the estimator and compare it to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection. The proposed method is applied to the NLST study dataset to uncover the effects of key clinical and imaging risk factors on patients' survival. Our findings provide valuable insights into the relationship between these factors and survival outcomes.

14:00 **Post-Estimation Strategies in Sparse Semiparametric Models for High-Dimensional Data Application**
S. Ejaz Ahmed

Abstract: In this talk, we propose shrinkage semiparametric estimation strategies for a partially linear regression model. In this framework, the regression parameter vector is partitioned into two sub-vectors: the first sub-vector gives the predictors of interest, i.e., main effects (for example, treatment effects), and the second sub-vector is for variables that may or may not need to be controlled. We establish both theoretically and numerically that proposed shrinkage strategy which combines two semiparametric estimators computed for the full model and the submodel outperform the semiparametric benchmark estimator. A real data example is given to show the usefulness of the proposed prediction strategy. Now-a-days high-dimensional data (HDD) analysis has become increasingly popular in a host of research arena. To analyze HDD, many penalized methods were introduced for simultaneous variable selection and parameters estimation when the model is sparse. However, a model may have sparse signals as well as several predictors with weak signals. In this scenario variable selection methods may not be able to distinguish predictors with weak signals and sparse signals. For this reason, we propose a high-dimensional shrinkage strategy to improve the prediction performance of a semiparametric submodel. We demonstrate that the proposed high-dimensional shrinkage strategy performs uniformly better than the penalized and machine learning methods in many cases. We numerically appraise relative performance of the proposed strategy. Some open research problems will be discussed, as well. [1] S. Ejaz Ahmed, Feryaal Ahmed and B. Yuzbasi (2023). Post-Shrinkage Strategies in Statistical and Machine Learning for High Dimensional Data. CRC Press, USA.

14:30 **Accommodating Time-Varying Heterogeneity in Risk Estimation: A Transfer Learning Approach**
Yu Shen, Jing Ning, Ziyi Li

Abstract: Cancer registries have been widely used in clinical research because of their easy accessibility and large sample size. To use cancer registry data as a complement to improve the estimation precision of individual risks of death for inflammatory breast cancer (IBC) patients at The University of Texas MD Anderson Cancer Center, we proposed to use transfer learning method for adaptive information borrowing. When transferring information for risk estimation based on the cancer registries (i.e., source cohort) to a single cancer center (i.e., target cohort), time-varying population heterogeneity needs to be appropriately acknowledged. However, there is no literature on how to adaptively transfer knowledge on risk estimation with time-to-event data from the source cohort to the target cohort while adjusting for time-varying differences in event risks between the two sources. Our goal is to address this statistical challenge by developing a transfer learning approach under the Cox proportional hazards model to allow data-adaptive levels of information borrowing. We develop a more accurate risk estimation model for the MD Anderson IBC cohort given various treatment and baseline covariates, while adaptively borrowing information from the National Cancer Database to improve risk assessment.

15:00 **Confidence Intervals and Simultaneous Confidence Bands Based on Deep Learning**
Asaf Ben Arie, Malka Gorfine

Abstract: Deep learning models have significantly improved prediction accuracy in various fields, gaining recognition across numerous disciplines. Yet, an aspect of deep learning that remains insufficiently tackled is the assessment of predictive uncertainty. Producing reliable uncertainty estimators could be crucial in practical terms. For instance, predictions associated with a high degree of uncertainty could be either overlooked or sent for further evaluation. Recent works in uncertainty quantification, including Bayesian credible intervals and a frequentist method inspired by the jackknife technique (leave-one-out), have proven to yield either invalid or too conservative intervals. Furthermore, there is currently no method for quantifying uncertainty that is capable of accommodating deep neural networks for survival data that involves right-censored outcomes. In this work we provide a valid non-parametric bootstrap method that correctly disentangles between data uncertainty and the noise inherent in the adopted optimization algorithm, ensuring that the resulting point-wise confidence intervals or the simultaneous confidence bands are accurate (i.e., valid and not overly conservative). The proposed method can be easily integrated into any deep neural networks without interfering with the training process. The utility of the proposed approach is illustrated by constructing simultaneous confidence bands for survival curves derived from survival data with right censoring.

15:30 - 16:00 Coffee Break

16:00 - 17:00 Contributed 3

Functional data analysis 3
Chair: Laura M. Sangalli
Room: Grande Auditorio

16:00 **Functional approaches to nonparametric risk reserving using standard chain ladder data**
Matus Maciak, Ivan Mizera, Michal Pesta

Abstract: Typical and commonly used reserving techniques are based on different parametric approaches using aggregated data--so-called run-off triangles. We propose some non-parametric alternatives that handle the underlying loss development triangles as functional profiles and we predict the overall claim reserve distribution through a permutation bootstrap. Theoretical justifications are provided and some practical implementation issues are addressed for the proposed non-parametric methods. In addition, a finite sample evaluation in terms of a full-scale comparison with standard (parametric) reserving techniques is carried on several hundreds of real run-off triangles against known real loss outcomes.

16:20 **Statistical Inference For Spectral Means Of Hilbert Space Valued Random Processes**
Daniel Rademacher, Jens-Peter Kreiss, Efstathios Paparoditis

Abstract: The subject of our analysis are discrete time processes that take their values in a (separable) Hilbert space, so called functional time series. A variety of statistics for functional time series allows for a representation as weighted average of corresponding periodogram operators over the frequency domain. We study consistency and asymptotic normality of such spectral mean estimators under mild assumptions. As a by-product, we derive some basic first and second order properties of periodogram operators as well. We show that weak convergence of spectral mean estimators can be deduced from the (joint) weak convergence of the sample autocovariance operators. The latter is established for a large class of weakly dependent functional time series, which admit expansions as Bernoulli shifts and the weak dependence is quantified by the condition of L^4-m-approximability. In order to facilitate the applicability of our results, we also propose a hybrid-bootstrap scheme, which combines a bootstrap-approach with a subsampling method to approximate the gaussian limit of spectral mean estimators.

16:40 **Bayesian Variable Selection for Function-on-Scalar Regression Models: A Comparative Analysis**
Camila de Souza, Pedro H. T. de Oliveira Sousa, Ronaldo Dias

Abstract: In this work, we developed a new Bayesian method for variable selection in function-on-scalar regression (FOSR). Our method uses a hierarchical Bayesian structure and latent variables to enable an adaptive covariate selection in FOSR. Extensive simulation studies show the proposed method's accuracy in estimating the functional coefficients and high capacity to select variables correctly. Furthermore, we conducted a substantial comparative analysis using the main competing methods: the BGLSS (Bayesian Group Lasso with Spike and Slab prior) method, the group LASSO (Least Absolute Shrinkage and Selection Operator), the group MCP (Minimax Concave Penalty), and the group SCAD (Smoothly Clipped Absolute Deviation). Results demonstrate that the proposed methodology is superior in correctly selecting covariates compared with the existing competing methods while maintaining a satisfactory level of goodness of fit. We also considered a COVID-19 dataset from Brazil as an application and obtained satisfactory results. In short, the proposed Bayesian variable selection model is highly competitive, showing significant predictive and selective quality.

Goodness-of-fit
Chair: M. Dolores Jiménez-Gamero
Room: Sala Polivalente 1.1

16:00 **Tests of uniformity on the sphere with data-driven parameters**
Alberto Fernández-de-Marcos, Eduardo García-Portugués

Abstract: Two new omnibus tests of uniformity for data on the sphere are proposed. The new test statistics exploit closed-form expressions for orthogonal polynomials, feature tuning parameters, and are related to a "smooth maximum" function and the Poisson kernel. We obtain exact moments of the test statistics under uniformity and rotationally symmetric alternatives, and give their null asymptotic distributions. We consider approximate oracle tuning parameters that maximize the power of the tests against known generic alternatives and provide tests that estimate oracle parameters through cross-validated procedures while maintaining the significance level. Numerical experiments explore the effectiveness of null asymptotic distributions and the accuracy of inexpensive approximations of exact null distributions. A simulation study compares the power to other tests of the Sobolev class, showing the benefits. The proposed tests are applied to the study of the nursing times of wild polar bears.

16:20 **Addressing missing data challenges: A multivariate goodness-of-fit testing perspective**
Danijel Aleksić, *Bojana Milošević*

Abstract: The problem of missing data is common when dealing with real data sets. Therefore it attracted the attention of scientists across various disciplines. However, its implications on goodness-of-fit testing remain relatively unexplored. In this study, we aim to bridge this gap in the literature by examining modifications to some commonly used tests designed for complete data to accommodate missingness issues. Our objectives are twofold: to demonstrate that the impact of imputation procedures remains significant across all types of missingness and to investigate the properties of test modifications under the assumption of missing completely at random (MCAR) data. Our findings, which include insights into the limiting null distributions for modified characteristic-function-based tests and extensive power analyses, lead to general recommendations for addressing missing data challenges within the context of goodness-of-fit testing.

16:40 **Single-index quantile regression models: a new lack-of-fit test**
Mercedes Conde-Amboage, Alvaro Arrojo-Vazquez

Abstract: Quantile regression models are usually used when a fully detailed study of a variable of interest with respect to certain explanatory variables is desired, without being limited to the study of a central tendency as it happens in mean regression. In this way, quantile regression makes it possible to perform a more complete and robust analysis of the data. On the other hand, parametric quantile regression models may sometimes not be flexible enough to collect all the information from the sample. In this context, semiparametric quantile regression models arises, in particular, single-index models have gained popularity in recent years. Along this talk, a new lack-of-fit test for single-index quantile regression models will be presented. The test is based on the cumulative sum of residuals following the ideas of He and Zhu (2003, Journal of the American Statistical Association). To approximate the critical values of the test, a wild bootstrap mechanism is used. An extensive simulation study was carried out that shows the good properties of the new test, under homo- and heteroscedastic regression models. Finally, with the aim of illustrating the usefulness of the new proposal in practice, the test is illustrated with real data about housing values in suburbs of Boston. The data set is available in the R package MASS under the name boston.

High-dimensional data 1
Chair: Claudio Agostinelli
Room: Sala Polivalente 1.2

16:00 **Robustifying and simplifying high-dimensional regression with applications to yearly stock returns and telematics data**
Michael Scholz, Maria Dolores Martinez-Miranda, Malvina Marchese, Jens Perch Nielsen

Abstract: The availability of a large number of variables that can have predictive power makes their selection in the regression context difficult. This paper considers robust and understandable low-dimensional estimators as building blocks to improve the overall predictive power by combining these building blocks in an optimal way. Our new algorithm is based on generalised cross-validation and builds the predictive model step-by-step forward from the simple mean to more complex predictive combinations. Practical applications to annual financial returns and actuarial telematics data show its usefulness for the financial industry and insurance.

16:20 **On the speed of the convergence of some kernel random forests.**
Isidoros Iakovidis

Abstract: Random forests are notable statistical learning algorithms introduced by Breiman in 2001 and they are widely used for classification and regression tasks. In this talk, we consider a specific class of random forest algorithms related to kernel methods, the KeRF (Kernel Random Forests.) In particular, we improve the rates of convergence for two explicit random forest algorithms, designed independently of the data set, named centered kernel random forest and uniform kernel random forest. The talk is based on the published joint work [1] with Nicola Arcozzi. [1] Iakovidis Isidoros, Nicola Arcozzi. Improved convergence rates for some kernel random forest algorithms[J]. Mathematics in Engineering, 2024, 6(2): 305-338. doi: 10.3934/mine.2024013

16:40 **Break detection procedures for high dimensional panel data**
Charl Pretorius, Heinrich Roodt

Abstract: We present a new test criterion for detecting changes in the means of independent panels in high-dimensional panel data. Unlike many existing procedures for time-dependent observations, the new test has the practical advantage of not requiring estimation of long-run variances and hence eliminates its reliance on choosing any bandwidth parameters. Moreover, the test statistic is shown to be entirely self-normalising in the sense that its null distribution is asymptotically pivotal, even in the presence of weak serial dependence in observations. This allows for the tabulation of general asymptotic critical values which may readily be used in applications, including in situations where the true underlying data generating process is unknown. Numerical results are presented which show that the new test alleviates certain finite-sample issues observed when using existing tests containing bandwidth parameters. The talk is concluded with an application to real data.

Imperfectly observed data
Chair: Luís Machado
Room: Pequeno Auditorio

16:00 **Infinitely divisible priors on exponent measures**
Florian Brück

Abstract: We provide a large toolbox for non-parametric priors on the survival function of multivariate random vectors. In essence, we assume that the survival function of the random vector can be expressed in terms of a so-called exponent measure. The prior is then employed by assuming that the exponent measure is infinitely divisible. This framework embeds many of the well-known non-parametric priors on cumulative hazard rates. The posterior distribution of the infinitely divisible exponent measure is derived and shown to exhibit certain conjugacy properties. Since the posterior distribution is quite complicated, we provide a general construction scheme for priors on multivariate random vectors, which allows to construct new prior distributions from simpler building blocks and allows for a closed form posterior representation.

16:20 **Tests of Missing Completely At Random based on sample covariance matrices**
Alberto Bordino, Tom Berrett

Abstract: We study the problem of testing whether the missing values of a potentially high-dimensional dataset are Missing Completely at Random (MCAR). We relax the problem of testing MCAR to the problem of testing the compatibility of a sequence of covariance matrices, motivated by the fact that this procedure is feasible when the dimension grows with the sample size. Tests of compatibility can be used to test the feasibility of positive semi-definite matrix completion problems with noisy observations, and thus our results may be of independent interest. Our first contributions are to define a natural measure of the incompatibility of a sequence of correlation matrices, which can be characterised as the optimal value of a Semi-definite Programming (SDP) problem, and to establish a key duality result allowing its practical computation and interpretation. By studying the concentration properties of the natural plug-in estimator of this measure, we introduce novel hypothesis tests that we prove have power against all distributions with incompatible covariance matrices. The choice of critical values for our tests rely on a new concentration inequality for the Pearson sample correlation matrix, which may be of interest more widely. By considering key examples of missingness structures, we demonstrate that our procedures are minimax rate optimal in certain cases. We further validate our methodology with numerical simulations that provide evidence of validity and power, even when data are heavy tailed.

16:40 **Efficient quantile regression under censoring using Laguerre polynomials**
Alexander Kreiss, Ingrid Van Keilegom

Abstract: In this talk we consider a novel methodology to estimate linear quantile regression models when the response is randomly right censored. The proposed methodology is based on approximating the error distribution by means of an extension of the Laplace distribution using Laguerre polynomials. More specifically, the extension is obtained by enriching the Laplace density by means of Laguerre polynomials, in such a way that the new density has by construction the property that its quantile of interest is equal to zero. Hence, when the error term has an Enriched Laplace density, it satisfies by construction the constraints of a quantile regression model. In addition, this extension is flexible enough to approximate many continuous distributions, if the number of parameters in this enriched Laplace distribution grows to infinity. We will show that with this new quantile regression model, we can obtain novel estimators of the quantile function, which are shown to be consistent and asymptotically normal. While we focus with our results on right-censored data, the proposed methodology is general and can directly be transferred to other forms of imperfect data. Coming back to right censored data, we also establish the asymptotic efficiency bound for general quantile regression in right censored data. Under general models, we understand models that require only that the quantile function of interest is linear in the observed covariates (rather than all quantile functions). We discuss efficiency of our estimator theoretically in a special case. We furthermore compare empirically the performance of our estimator compared to others in different scenarios. While doing so we also investigate when the efficiency bound is reached. Finally, to illustrate the performance of our estimator in practice, we apply the estimator to a dataset on Covid-19 patients. This is joint work with Ingrid Van Keilegom.

Network analysis
Chair: Zach Lubberts
Room: Sala Polivalente 1.3

16:00 **Network evolution by clustering attachment**
Natalia Markovich, Maksim Ryzhov, Marijus Vaiciulis

Abstract: The clustering attachment (CA) model is introduced in the paper Bagrow, Brockmann (2013) as an evolution tool of random networks. The attachment probability of a newly appending node to an existing node by the CA is determined by its clustering coefficient. The CA typically lowers the node clustering coefficient, in contrast to the well-known preferential attachment (PA) that executes so called 'rich-get-richer' models. The CA leads to numerous new phenomena like bursts of the modularity and a light-tailed distribution of node degrees. The idea of the CA is that nodes are drawn not towards hubs, but towards densely connected groups that is usual for social human behavior. Our results concern to CA models with zero and non-zero valued parameters. For the zero-valued CA we focus on the study of a total triangle count that is considered in the literature as an important characteristic of the network clustering. It is proved that the total triangle count tends to infinity a.s. for the latter model. For the non-zero valued CA we obtain the following theoretical properties: (1) The probability for a newly appended node to be attached to a unordered pair of existing nodes using the weighted sampling without replacement is obtained and it is proved that the collection of such probabilities forms a probability distribution; (2) The deviation of the consecutive mean clustering coefficients tends to zero over time; (3) The sequences of node degrees and triangle counts of any fixed node over time are proved to be submartingales. The results were obtained for any initial graph and for the CA without node and edge deletion. The results were approved by the simulation study. Moreover, the simulation study concerns to the CA with the node or edge deletion.

16:20 **Equivariant and Invariant Modelling of Complex Data**
Andreas Abildtrup Hansen, Aasa Feragen, Anna Calissano

Abstract: Non-Euclidean data, also known as Complex or Object Oriented Data, often come with intrinsic constraints and symmetries that define their non-standard geometry. Such constraints are, in many applications, encoded by group actions. Examples include the rotation group SO(n) for images or the node permutation group for graphs. The majority of the Non-Euclidean statistics- and machine learning literature has focused on defining a natural embedding space for the data, e.g. using quotient spaces, and subsequently extending standard data analysis methods to this context. Here, effort is concentrated on characterizing the geometry of the embedding space in order to extend data analysis tools to it. We will discuss a different perspective: To switch the attention from the embedding space to the model itself and constrain it to be equivariant or invariant with respect to the specific group action. This allows a different view on extending Euclidean tools to, implicitly, work on nonlinear group quotient spaces, via equivariant and invariant models. We describe the benefits of constraining popular predictive and generative models to be equivariant and how this constraint affects subsequent analysis. As an example, we consider the domain of graphs, where the permutation group acts on the nodes of the graph. We discuss how variational autoencoders, a popular generative neural network, can be generalized to the permutation equivariant setting, where it can be used for representation learning by examining its latent space. Next, we demonstrate how the model's equivariance affects the interpretability of the latent representations in a non-trivial manner. Lastly, we suggest methods for obtaining invariant representations post hoc in a manner respecting the latent space geometry, and thus providing the needed interpretability. The entire discussion is supported by a real-world example using a commonly used benchmark molecular dataset.

16:40 **Point processes on linear networks and how to address their comparison**
Maria Isabel Borrajo García, Ignacio González-Pérez, Wenceslao González-Manteiga

Abstract: Data sets representing the spatial location of a series of observations appear in a wide variety of scenarios, for example, trees in a forest, earthquakes in a region or traffic accidents in road networks. The latter is an example of point patterns which do not lay on a two-dimensional subregion of the plane, but which are constricted to a one-dimensional subset. These types of patterns are said to lay on a linear network. Analysing point processes on linear networks presents greater complexities than working on any Euclidean space, mainly because of the associated metric space. A vastly studied problem in Statistics is population comparison, i.e., determine whether two (or more) samples are generated by the same stochastic process. This problem also arises when dealing with point processes, for example, the distribution of two species of flora in a forest, outbreaks of natural or caused forest fires, car-car and car-motorcycle collisions on a road network... In the spatial point processes domain, this comparison problem has already been addressed, however this is not the case for point processes on other different domains. Inferential methods, as the ones proposed for the Euclidean plane, have not yet been developed regarding point processes on linear networks. In this work we study the two-sample problem for point processes on linear networks, proposing two specific testing methods, based on a Kolmogorov-Smirnov and a Cramer-von-Mises type statistics. A thorough simulation study is accomplished to detail the finite sample performance of our proposals. The test statistics are also applied to traffic collisions in Rio de Janeiro (Brazil).

Statistical learning and inference
Chair: Anna Calissano
Room: Sala Polivalente 1.4

16:00 **The gROC curve and the optimal classification system**
Pablo Martinez-Camblor, Sonia Perez Fernandez

Abstract: The binary classification problem (BCP) aims to correctly allocate subjects in one of two possible groups. The groups are frequently defined for having or not one characteristic of interest. With this goal, we are allowed to use different types of information. There is a huge number of methods dealing with this problem; including standard binary regression models, or complex machine learning techniques such as support vector machine, boosting, or perceptron, among others. When this information is summarized in a continuous score, we have to define classification regions (or subsets) which will determine whether the subjects are classified as positive, with the characteristic under study, or as negative, otherwise. The standard (or regular) receiver-operating characteristic (ROC) curve considers classification subsets in the way [c, ∞) and plot the true- against the false- positive rates (sensitivity against one minus specificity). The so-called generalized ROC, gROC, curve allows that both higher and lower values of the score will be associated with higher probabilities of being positive. Besides, the efficient ROC curve considers the optimal use of the scores without considering the potential impact on the derived classification subsets. In this document, we are interested in studying, comparing and estimating the transformations leading to the eROC and to the gROC curves. We will prove that, when the optimal transformation does have no relative maximum, both curves are equivalent. Besides, we investigate the use of the gROC curve on some theoretical models, explore the relationship between both the gROC and the eROC curves, and propose two non-parametric procedures for approximating the transformation leading to the gROC curve. The finite-sample behavior of the proposed estimators are explored in a Monte Carlo simulation study. Two real-data set illustrate the practical use of the methods.

16:20 **Bayesian Analysis of a Multivariate Density Ratio Model**
Victor De Oliveira

Abstract: Nowadays data are abundant, often being multivariate and coming from different sources. The goal in many of these situations is the efficient combination or fusion of the information from the different sources to answer the question(s) of interest, which results in more efficient and reliable inferences than using a single source. A useful approach to achieve this goal is the use of the so-called density ratio model, which makes minimal assumptions about the several multivariate distributions involved. In this talk, we investigate methods to perform Bayesian inference from several related multivariate data sources based on a multivariate density ratio model. To test the practical applicability of the proposed methodology, we use a previously analyzed dataset to quantify the effect of height and age on the weight of germ cell testicular cancer patients.

16:40 **An inference method to deal with multiple causes of failure**
Nora Villanueva, Marta Sestelo, Luís Meira-Machado, Javier Roca-Pardiñas

Abstract: Competing risks data arises when an individual may fail from different causes. The cumulative incidence function was proposed in order to estimate the marginal probability of a certain event in the presence of competing risks. One basic but important goal in the analysis of competing risk data is the comparison of these curves, for which limited literature exists. By determining groups of cumulative incidence functions, we can identify subgroups of individuals who have different probabilities of experiencing each of the competing events and this can be useful for understanding the heterogeneity of the population and tailoring interventions to specific subgroups. We proposed a new procedure that lets us not only test the equality of these curves but also group them if they are not equal. The proposed automatic method allows determining the composition of the groups as well as their number. Simulation studies show the good numerical behaviour of the proposed methods for finite sample size. The applicability of the proposed method is illustrated using real data.

17:00 - 19:00 Excursions

20:00 - 22:00 Dinner

9:00 - 10:00 Contributed 4

High-dimensional data 2

Chair: Natalie Neumeyer
Room: Sala Polivalente 1.2

9:00 **Self-normalized sums in high dimension: which covariance estimator?**
Emmanuelle Gautherat, Patrice Bertail, El Mehdi Issouani

Abstract: We propose a new estimator of the covariance in high-dimension that preserves the good properties of self-normalised sums. There are many methods using self-normalised sums. One example is the Hotelling test, which is constructed from a statistic that is exactly a self-normalised sum (to within one coefficient). The value of self-normalised sums in the context of observations arising from symmetric laws is well known, since they make it possible to dispense with hypotheses of moments in order to obtain checks on the tail of the distribution, for example. It is essentially this property that we want to preserve. On the one hand, self-normalised sums use the inverse of the covariance matrix, which does not exist in the context of high-dimensional data. On the other hand, the field of high-dimensional covariance estimators offers a wide variety of invertible estimators - in particular shrinkage covariance estimators - but these do not allow us to retain the self-normalisation structure that results in the absence of the moments hypothesis. We propose a regularized estimators of the covariance, of the type presented by Bodnar et al (2014) [1], but with a regularization that can be random and explicit, allowing to keep the good property of no moment hypothesis in the case of symmetric laws. The finite distance control of the tail distribution of the self-normalized sum constructed with this estimator will also be exposed. [1] Bodnar, A. K. Gupta, N. Parolya, On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix, Journal of Multivariate Analysis 132 (2014) 215–228.

9:20 **Adaptive clustering through composite entropy Minimization**
Thierry Dumont

Abstract: I will introduce a clustering method and offer both a theoretical analysis and an explanation for a phenomenon observed in the applied statistical literature since the 1990s. This phenomenon is the adaptability of the order when employing a clustering approach derived from the renowned Expectation Maximization (EM) algorithm. This method utilizes the minimization of an entropy-based criterion to achieve a balance between the entropy resulting from the composition of the mixture (including the number of components and their proximity to a uniform distribution) and the relative entropy concerning a specific family of functions. I propose a new statistical measure, the Relative Entropic Order (REO), which represents the optimal number of clusters for classification in relation to the chosen family of functions. We demonstrate the consistency of both the empirical REO and the optimal mixture. Our clustering method is distinctively adaptable to a wide array of data types, including multidimensional and heterogeneous datasets, even those with missing values. Following the discussion of theoretical and algorithmic foundations, I will showcase the effectiveness of the method through applications on both synthetic and real-world data. Specifically, we will explore how this method facilitates supervised classification, highlighting its versatility and efficacy.

9:40 **Tail Inference with Probability Weighted Moments**
Frederico Caeiro, Ayana Mateus, Dora Gomes

Abstract: The tail of a probability distribution capture important information about the magnitude and frequency of extreme events. Thus, understanding the upper tail behaviour is crucial in decision making and risk assessment. Statistical analysis and predictions of extreme events often require the estimation of the extreme value index, a key parameter which describes the tail heaviness. We present a new class of estimators of the extreme value index, based on Probability Weighted Moments (PWM), a flexible framework for tail inference.

Multiple testing and simultaneous inference

Chair: Ruth Heller
Room: Sala Polivalente 1.1

9:00 **Robust semi-parametric testing in generalized linear models with many responses**
Jesse Hemerik

Abstract: Generalized linear models are often misspecified due to overdispersion and heteroscedasticity. Existing quasi-likelihood methods for testing in misspecified models often do not provide satisfactory type I error rate control. We provide a novel semi-parametric test, based on a permutation-type approach. Our test often provides better type I error control than its competitors. Further, we consider the common scenario that there are multiple response variables. One example is RNA-Seq data, where the responses are counts. Another example is neuroimaging data on brain lesions (brain damage), where for every voxel the response is absence or presence of lesion. For each of the responses, association with the predictor of interest is tested. The challenge is then to deal with the multiple testing problem in a powerful and reliable way. To achieve this, we combine our approach with powerful permutation-based multiple testing methods.

9:20 **Post-hoc and Anytime Valid Permutation and Group Invariance Testing**
Nick Koning

Abstract: We study post-hoc (e-value-based) and post-hoc anytime valid inference for testing exchangeability and general group invariance. Our methods satisfy a generalized Type I error control that permits a data-dependent selection of both the number of observations n and the significance level. We derive a simple analytical expression for all exact post-hoc valid p-values for group invariance, which allows for a flexible plug-in of the test statistic. For post-hoc anytime validity, we derive sequential p-processes by multiplying post-hoc p-values. In sequential testing, it is key to specify how the number of observations may depend on the data. We propose two approaches, and show how they nest existing efforts. To construct good post-hoc p-values, we develop the theory of likelihood ratios for group invariance, and generalize existing optimality results. These likelihood ratios turn out to exist in different flavors depending on which space we specify our alternative. We illustrate our methods by testing against a Gaussian location shift, which yields an improved optimality result for the t-test when testing sphericity, connections to the softmax function when testing exchangeability, and an improved method for testing sign-symmetry.

9:40 **Blending Point-wise Inference and Cluster Mass Tests for powerful Massively Univariate Tests to control FWER in EEG data**
Olivier Renaud, Jaromil Frossard

Abstract: For the analysis of data coming from experiments based on time-locked electro/magneto-encephalogram (EEG/MEG), or more generally for the point-by-point comparison of smooth signals, the cluster mass test has been widely used to perform so-called massively univariate tests. It is a powerful method for detecting effects while controlling weakly the family-wise error rate (FWER), although its correct interpretation can only be performed at the cluster level without any point-wise conclusion. Somehow counterintuitive, it implies that the discovery of a significant cluster does not allow us to precisely localize the effect in time (or in space). We propose a new multiple comparisons procedure, the cluster depth tests, that both controls the FWER while allowing an interpretation at the time point level. We show the conditions for a strong control of the FWER, and a simulation study shows that the cluster depth tests achieve large power and guarantee the FWER even in the presence of physiologically plausible effects. By having an interpretation at the time point/voxel level, the cluster depth tests make it possible to take full advantage of the high temporal resolution of EEG recording and give a precise timing of the start and end of the significant effects. [1] Frossard J, Renaud O (2022) The cluster depth tests: Toward point-wise strong control of the family-wise error rate in massively univariate tests with application to M/EEG. NeuroImage 247:118824. https://doi.org/10.1016/j.neuroimage.2021.118824

Nonparametric econometrics 1
Chair: Miguel A. Delgado
Room: Grande Auditorio

9:00 **M-Estimation in Censored Regression Model using Instrumental Variables under Endogeneity**
Swati Shukla

Abstract: We propose and study M-estimation to estimate the parameters in the censored regression model in the presence of endogeneity, i.e., the Tobit model. In the course of this study, we follow two-stage procedures: the first stage consists of applying control function procedures to address the issue of endogeneity using instrumental variables, and the second stage applies the M-estimation technique to estimate the unknown parameters involved in the model. The large sample properties of the proposed estimators are derived and analyzed. The finite sample properties of the estimators are studied through Monte Carlo simulation and a real data application related to women's labor force participation.

9:20 **Weak convergence of the function-indexed sequential empirical process for nonstationary time series**
Florian Scholze

Abstract: Sequential empirical processes have several statistical applications, including change detection and goodness-of-fit testing. Studying their behaviour in different settings is therefore of both theoretical and practical interest. So far, the literature on the weak convergence of the function-indexed sequential empirical process under dependence seems to be limited to the stationary case. To partially close this gap, this work studies its weak convergence in a nonstationary setting, and it is shown to be asymptotically equicontinuous provided suitable maximal inequalities are available for the increments of its nonsequential counterpart. The assumed maximal inequalities implicitly contain some (nonspecific) dependence restrictions, but no additional dependence restrictions are imposed. Thereby, a certain level of generality is achieved, which enables a range of possible applications. Limitations, possible extensions and statistical applications are discussed.

9:40 **Multiscale Comparison of Nonparametric Trend Curves**
Marina Khismatullina, Michael Vogt

Abstract: We develop new econometric methods for the comparison of nonparametric time trends. In many applications, practitioners are interested in examining the trending behaviour of the observed time series. Among other things, they would like to know which trends have a different form and in which time intervals the differences occur. We design a multiscale test to formally approach these questions. Specifically, we develop a test which allows to make rigorous confidence statements about which time trends are different and where (i.e., in which time intervals) they differ. Based on our multiscale test, we further develop a clustering algorithm which allows to cluster the observed time series into groups with the same trending behaviour. We derive asymptotic theory for our test and clustering methods. The theory is complemented by a simulation study and two applications to house pricing and GDP growth data.

Nonparametric inference and estimation 1
Chair: Richard Samworth
Room: Pequeno Auditorio

9:00 **Testing Conditional Dependence**
Laura Freijeiro González, Wenceslao González-Manteiga, Manuel Febrero Bande

Abstract: Conditional dependence is an intricate type of dependence. This claims that, given two random variables, X and Y, these are dependent conditioned to a third one, Z. As a result, both variables, X and Y, could be independent, but present some conditional dependence relation given Z. We briefly review this implication and introduce the novel measure of conditional dependence of Wang et al. (2015): the conditional distance covariance coefficient (CDC). A discussion about its high computational cost and strategies to make its estimation possible in practice follows. Specifically, we focus on the proper selection of the bandwidth parameters for estimation and calibration in the vectorial framework of conditional dependence. We propose a new way of selecting automatic bandwidths for both and display how this selection outperforms its previous results. This is illustrated through simulations applied to conditional covariates selection. Eventually, we extend these ideas to more complex contexts.

9:20 **Inference for bivariate data with unobserved order**
Laura Dumitrescu

Abstract: Given a sequence of independent and identically distributed observations, we consider the situation when the order in each pair is not observed, but rather only the minimum and the maximum are recorded. In this case, it is not possible to construct the corresponding empirical distribution functions and nevertheless, several approaches to hypothesis testing of equal marginals have been proposed in the literature. In this talk, under a nonparametric framework, we relax the assumption of independence between the components of each pair and obtain estimators of their distributions. Simulation results and a real-data example are used to illustrate the efficacy of the proposed method.

9:40 **Exponential bounds for penalized Hotelling statistics**
El Mehdi Issouani, Patrice Bertail, Emmanuelle Gautherat

Abstract: In this presentation, we study the behavior of Self-normalized sums in high-dimensional settings. We derive exponential bounds for regularized Hotelling's T^2 statistics, considering the high-dimensional nature of the problem. We investigate the finite sample properties of the tail of these statistics, establishing exponential bounds for symmetric distributions and general distributions with weak moment assumptions (without assuming exponential moments). This talk aims to offer a thorough understanding of the exponential inequalities related to regularized Hotelling’s T^2 statistics, stressing their importance in high-dimensional statistical analysis. We will also discuss the practical implications of our results and suggest potential directions for future research.

Set estimation and inference
Chair: Clément Levrard
Room: Sala Polivalente 1.3

9:00 **Data Depth for Probability Measures**
Pierre Lafaye de Micheaux, Pavlo Mozharovskyi, *Myriam Vimond*

Abstract: Statistical data depth measures the centrality of a given point in space with respect to a finite sample, or with respect to a probability measure on that space. Over the last few decades, this seminal idea of data depth has evolved into a powerful tool that has proven useful in various fields of science. Recently, the notion of data depth was extented to unparametrized curves (De Miceaux et al, 2021). In this work, we go further and propose a notion of data depth suitable for data represented as probability measures. We have in mind applications with finite finite point processes, with distributions of random closed sets, or with models of germ grain coverage. Depending on the geometry of the data, we investigate adaptations of this depth, for example by introducing a weight (as in Kotik et al., 2017). We show that the depth satisfies the theoretical requirements of general depth functions that are meaningful for applications.

9:20 **Shape constraints beyond convexity**
Alejandro Cholaquidis, Leonardo Moreno, *Beatriz Pateiro-López*

Abstract: Convexity is a key concept that has been studied extensively in mathematics. In the context of set estimation, convexity is perhaps the most classical geometric restriction. Since the early works in the 1960s, many contributions have been made in this field, with a focus on various aspects including support estimation, volume and boundary measure estimation or level set estimation. Nonetheless, convexity can be limiting when dealing with problems that involve more flexible shape constraints. To address this limitation, a lot of research has focused on extending the concept of convexity to more general shape constraints. Standardness is one such shape restriction that is commonly used in set estimation and can be viewed as both a geometric and probabilistic condition. In this work, we review this notion, its connection to other geometrical restrictions and some methods for the problem of estimating the standardness constant.

9:40 **On Improved Semi-Parametric Bounds For Tail Probability And Expected Loss**
Artem Prokhorov, Erick Li

Abstract: We revisit the fundamental issue of tail behavior of accumulated random realizations when individual realizations are independent, and we develop new sharper bounds on the tail probability and expected linear loss. The underlying distribution is semi-parametric in the sense that it remains unrestricted other than the assumed mean and variance. Our sharp bounds complement well-established results in the literature, including those based on aggregation, which often fail to take full account of independence and use less elegant proofs. New insights include a proof that in the non-identical case, the distributions attaining the bounds have the equal range property, and that the impact of each random variable on the expected value of the sum can be isolated using an extension of the Korkine identity. We show that the new bounds not only complement the extant results but also open up abundant practical applications, including improved pricing of product bundles, more precise option pricing, more efficient insurance design, and better inventory management. Paper available at https://drive.google.com/file/d/1g9yWOToMup2QaNezjmyUT8qs-OlWro_3/view?usp=sharing

10:00 - 11:00 Keynote Talk 3
Chair: Inês Sousa
Room: Grande Auditorio

10:00 **Common atoms mixture models in two biostatistical inference problems**
Peter Mueller

Abstract: We consider two examples of statistical inference for two related populations. In one example we characterize two patient populations that are relevant in the construction of a clinical study design, and propose a method to adjust for detected differences. The second example is about characterizing immune cell populations under two biologic conditions of interest and identify shared versus condition-specific homogeneous cell subpopulations. Bayesian inference in both applications requires prior probability models for two or more related distributions. We build on extensive literature on such models based on Dirichlet process priors. Models are commonly known as dependent Dirichlet processes (DDP), with many variations and extensions beyond the Dirichlet process model. The special feature in the two motivating applications is the focus on differences in the heterogeneity of the related populations, with one application aiming to adjust for such differences, and the other application aiming to identify and understand immune cell subtypes that are characteristic for one or the other condition. We briefly review the extensive literature on DDP models and then introduce variations of DDP priors suitable for these inference goals. The common structure are common atoms mixture models with highly structured priors on the weights.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk 2

Chair: Jacobo de Uña-Álvarez
Room: Grande Auditorio

11:30 **Specification tests for statistical models with recent results and applications**
Wenceslao González-Manteiga

Abstract: Specification tests in statistical models refer to the assertion or rejection of some assumption about the statistical model structure. These cover well studied topics as for example the goodness-of-fit for distributions, densities or regression models between others. The objectives, in general, can be different: Testing if one regression model belongs to some parametric family, the comparison of the structure of k populations, testing the significance of covariates in one high dimensional regression model…etc. In this talk we will revise the different methodologies developed in the last 25 years (of my academic life) with recent results in complex models with some different applications.

12:30 - 13:30 Lunch

13:30 - 15:30 Invited 7

Statistics in the AI era: different perspectives
Organizer: Sophie Langer
Chair: Sophie Langer
Room: Grande Auditorio

13:30 **Class probability matching for label shift adaptation**
Annika Betken, Hongwei Wen, Hanyuan Hang

Abstract: Against the background of classification tasks, domain adaptation aims at building machine learning algorithms that on the basis of labeled (training) data from a source domain generalize well to unlabeled (test) data from a different, but related target domain. The label shift problem of domain adaptation refers to a change in the distribution of labels between source and target domain while it presupposes that the conditional distributions of features given labels are the same in both domains. To solve the label shift adaptation problem, we propose to match class probabilities in source and target domain, thereby establishing a novel method for label shift adaptation called “Class Probability Matching” (CPM). We show that class probability matching is able to maintain the same theoretical guarantees as existing feature probability matching frameworks, while significantly improving upon their computational efficiency. Within the CPM framework we propose an algorithm named “Class Probability Matching with Calibrated Networks (CPMCN) for target domain classification. This algorithm is based on classification in the source domain through calibrated neural networks. We establish a generalization bound for the CPMCN method and compare it to established procedures through a performance analysis on real data.

14:00 **Differentially private penalized M-estimation via noisy optimization**
Marco Avella Medina

Abstract: We propose a noisy composite gradient descent algorithm for differentially private statistical estimation in high dimensions. We begin by providing general rates of convergence for the parameter error of successive iterates under assumptions of local restricted strong convexity and local restricted smoothness. Our analysis is local, in that it ensures a linear rate of convergence when the initial iterate lies within a constant-radius region of the true parameter. At each iterate, multivariate Gaussian noise is added to the gradient in order to guarantee that the output satisfies Gaussian differential privacy. We then derive consequences of our theory for linear regression and mean estimation. Motivated by M-estimators used in robust statistics, we study loss functions which downweight the contribution of individual data points in such a way that the sensitivity of function gradients is guaranteed to be bounded, even without the usual assumption that our data lie in a bounded domain. We prove that the objective functions thus obtained indeed satisfy the restricted convexity and restricted smoothness conditions required for our general theory. We then show how the private estimators obtained by noisy composite gradient descent may be used to obtain differentially private confidence intervals for regression coefficients, by leveraging work in Lasso debiasing proposed in high-dimensional statistics. We complement our theoretical results with simulations that illustrate the favorable finite-sample performance of our methods.

14:30 **Wasserstein Generative Adversarial Networks are Minimax Optimal Distribution Estimators**
Arthur Stéphanovitch, Eddie Aamari, Clément Levrard

Abstract: We provide non asymptotic rates of convergence of the Wasserstein Generative Adversarial networks (WGAN) estimator. We build neural networks classes representing the generators and discriminators which yield a GAN that achieves the minimax optimal rate for estimating a certain probability measure μ with support in $\mathbb{R}^p$. The probability μ is considered to be the push forward of the Lebesgue measure on the d-dimensional torus $\mathbb{T}^d$ by a map $g^\star:\mathbb{T}^d\rightarrow \mathbb{R}^p$ of smoothness $\beta+1$. Measuring the error with theγ-Hölder Integral Probability Metric (IPM), we obtain up to logarithmic factors, the minimax optimal rate $O(n^{-\frac{\beta+\gamma}{2\beta +d}}\vee n^{-\frac{1}{2}})$ where n is the sample size,β determines the smoothness of the target measure μ, γ is the smoothness of the IPM ($\gamma=1$ is the Wasserstein case) and $d\leq p$ is the intrinsic dimension of μ. In the process, we derive a sharp interpolation inequality between Hölder IPMs. This novel result of theory of functions spaces generalizes classical interpolation inequalities to the case where the measures involved have densities on different manifolds.

15:00 **Dropout Regularization Versus L2-Penalization in the Linear Model**
Gabriel Clara, Sophie Langer, Johannes Schmidt-Hieber

Abstract: We investigate the statistical behavior of gradient descent iterates with dropout in the linear regression model. In particular, non-asymptotic bounds for the convergence of expectations and covariance matrices of the iterates are presented. The results shed more light on the widely cited connection between dropout and L2-regularization in the linear model. We indicate a more subtle relationship, owing to interactions between the gradient descent dynamics and the additional randomness induced by dropout. Based on joint work with Sophie Langer and Johannes Schmidt-Hieber

Bayesian nonparametrics for high-dimensional and complex models
Organizer: Ismaël Castillo
Chair: Ismaël Castillo
Room: Pequeno Auditorio

13:30 **A variational Bayes approach to debiased inference in high-dimensional linear regression**
Luke Travis, Ismaël Castillo, Alice L'Huillier, Kolyan Ray

Abstract: We consider statistical inference on a subset of a high-dimensional parameter in sparse linear regression. It is well-known that high-dimensional procedures such as the LASSO can provide biased estimators for this problem, and debiasing such procedures is well-studied in the frequentist literature. We develop a scalable variational Bayesian approach to this problem, motivated by a natural choice of prior based on a re-parameterisation given by a decomposition of the likelihood. We investigate the empirical performance of this procedure for estimation and uncertainty quantification via multidimensional credible sets, and establish theoretical guarantees in the form of a Bernstein-von Mises Theorem. Joint work with Ismael Castillo, Alice L’Huillier and Kolyan Ray.

14:00 **Bayes in the extreme**
Surya Tokdar

Abstract: Statistical analyses of heavy tailed data bring in a unique set of questions. Often the scientific focus shifts to the tails of the distribution, e.g., to forecasting the 100-year daily precipitation or to identifying predictors which influence extreme low birthweight. Parametric models, whose fit is largely dictated by the central bulk of the data, may not do justice to capturing tail structures. At the same time, purely nonparametric approaches may prove futile in effectively smoothing information from sparse observations across an elongated tail. Toward more effective statistical analyses of heavy tailed data, I will introduce a class of semiparametric Bayesian methods for density estimation and quantile regression. With a carefully chosen nonparametric prior distribution, the density estimation method will be shown to simultaneously guarantee accurate estimation of the density function and its tail index, both at near optimal minimax rate. The related quantile regression methodology will be shown to offer a powerful and yet interpretable generalization of standard linear regression to what one might call a Quantile Linear Model. The QLM, complete with an identification of residual noise, gives a model-based idealization of quantile regression retaining the ability to quantify differential predictor influence on the tails while simultaneously adjusting for noise correlation. I will discuss how QLM leads to a comprehensive inferential framework with the added qualities of model fit assessment and model selection.

14:30 **Almost-parallel Bayesian Gaussian Graphical Modelling in High-Dimensions**
Deborah Sulem, David Rossell, Jack Jewson

Abstract: Gaussian graphical models (GGM) are widely used to analyse the dependence structure among variables. However, when the number of observed variables is large, the computational demands of estimating a high-dimensional graphical model have limited the scope of applications. In this work, we introduce a scalable, interpretable, and fully-Bayesian method for estimating a high-dimensional GGM. Our method capitalises on a discrete spike-and-slab parametrisation of the prior distribution, leading to a truly-sparse estimated graphical model. We propose an efficient Block Gibbs algorithm to sample from the posterior distribution, together with an almost-parallel version which exploits the relationship between the conditional dependence structure and a linear regression model. This strategy facilitates decomposing the high-dimensional estimation problem into sub-components, allowing the application of efficient methodologies originally developed for linear regression. We empirically demonstrate that structure learning and statistical efficiency is improved by our discrete parametrisation.

15:00 **Convergence rates of deep Gaussian process regression**
Aretha Teckentrup

Abstract: Deep Gaussian processes have proved remarkably successful as a tool for various statistical inference and machine learning tasks. This success relates in part to the flexibility of these processes and their ability to capture complex, non-stationary behaviours. In this talk, we will introduce the general framework of deep Gaussian processes, and consider their use as prior distributions in regression and interpolation tasks. We present novel results on the convergence of the methodology as the number of data points goes to infinity.

Conformal and simultaneous inference
Organizer: Etienne Roquain
Chair: Etienne Roquain
Room: Sala Polivalente 1.1

13:30 **Combining exchangeable p-values**
Matteo Gasparin, Aaditya Ramdas

Abstract: Significant recent progress has been made on deriving combination rules that can take as input a set of arbitrarily dependent p-values, and produce as output a single valid p-value. Here, we show that under the assumption of exchangeability of the p-values, many of those rules can be improved (made more powerful). While this observation by itself has practical implications (for example, under repeated tests involving data splitting), it also has implications for combining arbitrarily dependent p-values, since the latter can be made exchangeable by applying a uniformly random permutation. In particular, we derive several simple randomized combination rules for arbitrarily dependent p-values that are more powerful than their deterministic counterparts, improving on well known rules like ''twice the median'' and ''twice the average''. Then, we show how these results can be used in the context of conformal prediction.

14:00 **Polya trees for nonparametric shrinkage estimation in high dimensional GLMs**
Asaf Weinstein, Jonas Wallin, Daniel Yekutieli, Malgorzata Bogdan

Abstract: In a given generalized linear model (GLM) with fixed effects, and avoiding any assumptions such as sparsity, what is the optimal regularized estimator of the coefficients? Relying on concepts of exchangeability—but without making any actual modeling assumptions on the coefficients—we first propose as a contender the Bayes estimator against an *ideal* prior that assigns equal mass to every permutation of the true coefficient vector, and show some optimality properties in both the frequentist and Bayesian frameworks. We then set out to mimic this oracle by postulating a nonparametric hierarchical Bayes model, taking the coefficients to be iid draws from an unknown distribution \pi, which itself is assigned a Polya tree prior. Posterior inference is facilitated by a Gibbs sampling algorithm leveraging conditional conjugacy of Polya trees. We show in simulations for linear and logistic regression that the posterior mean of \pi approximates well the empirical distribution of the true coefficients, effectively solving a nonparametric deconvolution problem. The Bayes estimators for the coefficients, in turn, can be thought of roughly as posterior estimates with respect to an iid prior learned nonparametrically from the data, and in our simulations perform almost as well as the oracle.

14:30 **Selecting informative conformal prediction sets with false coverage rate control**
Ruth Heller, Etienne Roquain, Ulysse Gazin, Ariane Marandon

Abstract: In supervised learning, including regression and classification, conformal methods provide prediction sets for the outcome/label with finite sample coverage for any machine learning predictors. We consider here the case where such prediction sets come after a selection process. The selection process requires that the selected prediction sets be `informative' in a well defined sense. We consider both the classification and regression settings where the analyst may consider as informative only the samples with prediction sets small enough, excluding null values, or obeying other appropriate `monotone' constraints. We develop a unified framework for building such informative conformal prediction sets while controlling the false coverage rate (FCR) over the selected. Our new procedure, InfoSP (Informative Selective Prediction sets), achieves FCR control for the selected examples that are all informative, assuming full exchangeability of the calibration and test samples (for both regression and classification) or class-conditional exchangeability (for classification). We introduce a second procedure, InfoSCOP (Informative Selective Conditional Prediction sets), that has the same theoretical properties as InfoSP, assuming full exchangeability of the calibration and test samples, but it enables an initial selection step that is aimed at eliminating (at least some of) the examples for which informative prediction sets cannot be constructed. Further selection then takes place in order to ensure that all reported prediction sets are informative. We show the usefulness of InfoSP and InfoSCOP on real and simulated data.

15:00 **Transductive conformal inference with adaptive scores**
Ulysse Gazin, *Gilles Blanchard*, Etienne Roquain

Abstract: Conformal inference is a fundamental and versatile tool that provides distribution-free guarantees for many machine learning tasks. We consider the transductive setting, where decisions are made on a test sample of m new points, giving rise to m conformal p-values. While classical results only concern their marginal distribution, we show that their joint distribution follows a Pólya urn model, and establish a concentration inequality for their empirical distribution function. The results hold for arbitrary exchangeable scores, including adaptive ones that can use the covariates of the test+calibration samples at training stage for increased accuracy. We demonstrate the usefulness of these theoretical results through uniform, in-probability guarantees for two machine learning tasks of current interest: interval prediction for transductive transfer learning and novelty detection based on two-class classification. This is joint work with Ulysse Gazin and Etienne Roquain.

Extrapolation methods for extreme values
Organizer: Abdelaati Daouia
Chair: Abdelaati Daouia
Room: Sala Polivalente 1.2

13:30 **Estimation of marginal excess moments for Weibull-type distributions**
Yuri Goegebeur, Armelle Guillou, Jing Qin

Abstract: We consider the estimation of the marginal excess moment (MEM), which is defined for a random vector (X,Y) and a parameter $\beta >0$ as $\mathbb E[(X-Q_X(1-p))_+^\beta|Y> Q_Y(1-p)]$ provided $\mathbb E|X|^\beta< \infty$, and where $y_+:=\max(0,y)$, Q_X and Q_Y are the quantile functions of X and Y respectively, and $p\in (0,1)$. Our interest is in the situation where the random variable X is of Weibull-type while the distribution of Y is kept general, the extreme dependence structure of (X,Y) converges to that of a bivariate extreme value distribution, and we let$p \downarrow 0$ as the sample size $n \to \infty$. By using extreme value arguments we introduce an estimator for the marginal excess moment and we derive its limiting distribution. The finite sample properties of the proposed estimator are evaluated with a simulation study and the practical applicability is illustrated on a dataset of wave heights and wind speeds.

14:00 **A conditional tail expectation type risk measure for time series**
Armelle Guillou, Yuri Goegebeur, Jing Qin

Abstract: We consider the estimation of the conditional expectation $\mathbb E(X_h |X_0>Q_X(1-p))$ at extreme levels, where $(X_t)_{t\in \mathbb Z}$ is a strictly stationary $\beta-$mixing time series, Q_X its associated quantile function, $p\in (0, 1)$ and h a positive integer. We use the multivariate regular variation framework and start to consider the case of non-negative time series. A two-step method is used in order to propose an estimator of this risk measure: first, by introducing an estimator in the intermediate case and, then, by extrapolating outside the data by a Weissman-type construction. Under suitable assumptions, we prove the weak convergence of the estimator of this risk measure. Subsequently, we extend our approach to the case of real-valued time series by using the decomposition of the original time series into the positive and negative parts and we prove again the weak convergence of the proposed estimator under additional assumptions. Some simulations are provided in order to illustrate the performance of our estimator.

14:30 **Functional Extreme-PLS**
Stephane Girard, *Cambyse Pakzad*

Abstract: We propose an extreme dimension reduction method extending the Extreme-PLS approach to the case where the covariate lies in a possibly infinite-dimensional Hilbert space. The ideas are partly borrowed from both Partial Least-Squares and Sliced Inverse Regression techniques. As such, the method relies on the projection of the covariate onto a subspace and maximizes the covariance between its projection and the response conditionally to an extreme event to capture most of the tail information. Moreover, we link the covariate and the heavy-tailed response through a non-linear inverse single-index model and our goal is to infer the index in this regression framework. We propose a new family of estimators and show its asymptotic consistency with convergence rates under the model. Assuming mild conditions on the noise, most of the assumptions are stated in terms of regular variation unlike the standard literature on SIR and single-index regression. Finally, we illustrate our results on a finite-sample study with synthetic functional data \cambyse{as well as on a real data application about financial assets.

15:00 **Extreme expectile estimation for short-tailed data**
Abdelaati Daouia, Simone Padoan, Gilles Stupfler

Abstract: The use of expectiles in risk management has recently gathered remarkable momentum due to their excellent axiomatic and probabilistic properties. In particular, the class of elicitable law-invariant coherent risk measures only consists of expectiles. While the theory of expectile estimation at central levels is substantial, tail estimation at extreme levels has so far only been considered when the tail of the underlying distribution is heavy. The article we will present is the first work to handle the short-tailed setting where the loss (e.g. negative log-returns) distribution of interest is bounded to the right and the corresponding extreme value index is negative. We derive an asymptotic expansion of tail expectiles in this challenging context under a general second-order extreme value condition, which allows to come up with two semiparametric estimators of extreme expectiles, and with their asymptotic properties in a general model of strictly stationary but weakly dependent observations. A simulation study and a real data analysis from a forecasting perspective are performed to verify and compare the proposed competing estimation procedures.

Recent advances in cure models
Organizer: Ricardo Cao
Chair: Ricardo Cao
Room: Sala Polivalente 1.3

13:30 **A presmoothed estimator for the cure rate in mixture cure models**
Ana López-Cheda, María Amalia Jácome Pumar, Samuel Saavedra

Abstract: Current cancer treatments caused an increased ratio of cured patients or, at least, a long term survival. In order to accommodate the insusceptible proportion of subjects, a cure fraction can be explicitly incorporated into survival models and as a consequence, cure models arise. The goals in cure models are usually to estimate the cure rate and the probability of survival of the uncured patients up to a given point of time (latency). Although, in the literature, parametric and semiparametric models have been considered, nonparametric estimation methods for cure models have attracted much attention in the last few years. A presmoothed nonparametric estimator for the probability of cure in mixture cure models is introduced. The methodology in Cao and Jácome (2004) is considered to improve the cure rate estimator in López-Cheda, Cao, Jácome and Van Keilegom (2017). In addition, the finite sample performance of the proposed estimator is assessed and compared with existing estimators in an extensive simulation study. Finally, the proposed method is applied to a real dataset. [1] Cao, R. and Jácome, M. A. (2004). Presmoothed kernel density estimation for censored data. Journal of Nonparametric Statistics, 16, 289-309. [2] López-Cheda, A., Cao, R., Jácome, M. A. and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Computational Statistics & Data Analysis, 105, 144–165.

14:00 **High dimensional mixture cure models: an application in cardio-oncology**
Beatriz Piñeiro Lamas, Ricardo Cao, Ana López Cheda

Abstract: In survival analysis, there are situations in which not all subjects are susceptible to the final event. For example, if the event is a cancer therapy-related adverse effect, there will be a fraction of patients (considered as cured) that will never experience it. Mixture cure models allow to estimate the probability of cure and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates. We fill this gap by proposing single-index mixture cure models. They allow working with a vector covariate and assume that the survival function depends on it through an unknown linear combination, that can be estimated by maximum likelihood. The proposed models are extended to functional covariates and a preprocessing algorithm is implemented to deal with medical images. The methodology is applied to a cardiotoxicity dataset. The goal is to determine whether (and how) certain factors affect the probability of experiencing the cardiovascular problem and the amount of time it takes for it to manifest. Understanding risk factors may lead to a personalized preventive medicine.

14:30 **Nonparametric inference for the mixture cure model with partially known cured observations**
Wende Clarence Safari, Ignacio Lpez-de-Ullibarri, *María Amalia Jácome Pumar*

Abstract: The mixture cure model (MCM) is a statistical framework used in survival analysis to account for the presence of both cured and uncured individuals in a study population. MCM assumes that the population is a mixture of two sub-groups: those whose event is certain not to occur are “cured” (or long-term survivors) and those who will experience the event are known to be “uncured” (or susceptible). However, a challenge in applying the MCM arises from the fact that long-term survivors are not directly observed to be cured; instead, they are typically censored at the study's conclusion, leaving the cure status latent in right-censored subjects. Nevertheless, in many medical studies, certain individuals may indeed possess a cured status. For instance, diagnostic procedures can offer insights into whether a subject may be considered as cured or not. Additionally, in certain types of cancer, it is highly improbable to experience a recurrence beyond a specific time post-treatment, known as the cure threshold. Consequently, patients with observed times surpassing this threshold can reasonably be considered as cured. In this talk, three nonparametric estimators will be explored within the context of the MCM when the cure status is partially known. These estimators will specifically target the estimation of the survival function, cure probability, and latency function (survival function of “uncured” individuals). An illustrative example from a study of COVID-19 patients hospitalised in Galicia (Spain) during the outbreak of the pandemic will be presented. The event of interest was the admission of a COVID patient to the ICU from the hospital ward. The study aimed to model the lengths of stay of patients in hospital centres and estimate the probability of a patient initially admitted to the ward subsequently requiring ICU admission, along with the duration until such admission occurs.

15:00 **Effect of a covariate in the cure rate of a mixture cure model using distance correlation**
Blanca Estela Monroy Castillo, María Amalia Jácome Pumar, Ricardo Cao

Abstract: The concept of dependence among observations plays a central role in many fields. In survival analysis, measuring the relationship between the lifetime and a covariate is usually of major interest. One of the goals in cure models (Peng and Yu, 2021) is to test whether a covariate influences the cure rate. Distance correlation (Székely et al 2007) is a novel class of multivariate dependence coefficients with advantages over classical correlation coefficients: it is applicable to random vectors of arbitrary dimensions not necessarily equal, and it is zero if and only if the vectors are independent. Different estimators have been compared in Monroy-Castillo et al (2024). In a standard survival model, Edelmann et al (2020) proposed an estimator for the distance covariance between covariates and survival times under right-censoring. But to the best of our knowledge, distance correlation has not been applied yet in the presence of cured individuals. We propose to study the effect of a covariate on the probability of cure by means of the distance correlation between this covariate and the cure indicator. The main challenge is to handle the missingness of the cure indicator of the censored individuals.

High-dimensional regression
Organizer: Ursula Mueller
Chair: Ursula U. Müller
Room: Sala Polivalente 1.4

13:30 **A Mean Field Approach to Empirical Bayes Estimation in High-dimensional Linear Regression**
Bodhisattva Sen, Sumit Mukherjee, Subhabrata Sen

Abstract: In this talk we consider the problem of empirical Bayes (EB) inference in Bayesian high-dimensional regression where the regression coefficients are drawn i.i.d. from an unknown prior. EB procedures estimate the prior probability distribution (in a Bayesian statistical model) from the data. To estimate this prior distribution we propose and study a "variational empirical Bayes" approach — it combines EB inference with a variational approximation (VA). The idea is to approximate the intractable marginal log-likelihood of the response vector --- also known as the "evidence" --- by the evidence lower bound (ELBO) obtained from a naive mean field (NMF) approximation. We then maximize this lower bound over a suitable class of prior distributions in a computationally feasible manner. We show that the marginal log-likelihood function can be (uniformly) approximated by its mean field counterpart. More importantly, under suitable conditions, we establish that this strategy leads to consistent approximation of the true posterior and provides asymptotically valid posterior inference for the regression coefficients.

14:00 **Statistical inference for the error distribution in functional linear models**
Natalie Neumeyer

Abstract: Some recent results on functional linear models with scalar response and functional covariate are presented. In those models it may be more challenging to deal with residual-based procedures than in regression models with vector-valued covariates. We present procedures for testing for changes in the error distribution, goodness-of-fit testing, and testing for independence of covariates and errors. We also consider models with vector-valued responses and functional covariates. Here the dependence between the components of the response, given the covariate, can be modeled by the copula of vector-valued errors, and we present the asymptotics of the empirical copula function.

14:30 **Adaptive variable selection in sparse nonparametric models**
Natalia Stepanova, Marie Turcicova

Abstract: We study the problem of adaptive variable selection in a Gaussian white noise model of intensity ε under certain sparsity and regularity constraints on an unknown regression function f. The d-variate regression function f is assumed to be the sum of functions each depending on a smaller number k of variables (1 ≤ k ≤ d). These functions are unknown to us and only few of them are nonzero. In this talk, we address the problem of identifying the nonzero k-variate components of f when d tends to infinity as ε tends to zero and k is either fixed or k tends to infinity and k=o(d) as ε tends to zero. This may be viewed as a variable selection problem. We propose an adaptive selection procedure that, under certain model assumptions, identifies exactly all nonzero k-variate components of f. In addition, we establish conditions under which exact identification of the nonzero components of f is impossible. These conditions ensure that the proposed selection procedure is the best possible in the asymptotically minimax sense with respect to the Hamming risk.

15:00 **Variable selection by voting**
Ursula U. Müller

Abstract: We consider a sparse linear model with a fixed design matrix in a high dimensional scenario. We introduce a new variable selection procedure called "voting", which combines the results from multiple regression models with different penalized loss functions to select the relevant predictors. A predictor is included in the final model if it receives enough votes, i.e. is selected by most of the individual models. By employing multiple different loss functions our method takes various properties of the error distribution into account. This is in contrast to the standard penalized regression approach, which typically relies on just one criterion. When that single criterion is not met the standard approach is likely to fail, whereas our method is still able to identify the underlying sparse model. Working with the voting procedure reduces the number of predictors that are incorrectly selected, which simplifies the structure and improves the interpretability of the fitted model. We prove model selection consistency and illustrate the advantages of our method numerically with simulations. This talk is based on joint work with Guorong Dai (Fudan University) and Raymond Carroll (Texas A&M University)

Recent advances in multivariate time series analysis
Organizer: Giovanni Motta
Chair: Giovanni Motta
Room: Sala Polivalente 1.5

13:30 **Measuring and Predicting Cyclical Turning Points, Gaps, and Drawdowns**
Tommaso Proietti

Abstract: By locating the running maxima and minima of a time series, and measuring the current deviation from them, it is possible to generate processes that are relevant for the analysis of the business cycle and for characterizing bull and bear phases in financial markets. The talk focuses, in particular, on the output gap (the deviation of current output from its potential) and the drawdown of a financial asset (the potential loss in the value of an asset when it deviates from its historical peak). Further, the measurement of the current lead time from the running maxima and minima originates Markov chains, which are informative on the duration of market phases and time reversibility of the underlying process, and, together with the distribution of output growth and asset returns, determine the properties of the output gap and the drawdown. A new algorithm for dating peaks and troughs of the output (price) process, delimiting expansions and recessions (bear and bull market phases), is derived with the support of the two chains. We finally discuss out-of-sample prediction and robust estimation of gaps and drawdowns.

14:00 **Non-parametric estimation of Dynamic Factor Models in the frequency domain**
Giovanni Motta, Michael Eichler

Abstract: The Generalized Dynamic Factor Model introduced in Forni and Lippi (2001) has become very popular in the theory and practice of large panels of time series data. The asymptotic properties of the corresponding estimators have been studied in Forni et al. (2000). Those estimators rely on Brillinger's dynamic principal components and thus involve two-sided filters, which leads to rather poor forecasting performances. Forni et al. (2017) derive the asymptotic properties of a semi-parametric estimator of loadings and common shocks based on one-sided filters. However, compared to the model by Forni et al. (2000), the model by Forni et al. (2017) rely on the additional assumptions that the common components have rational spectral density and admit a finite autoregressive representation. Moreover, the estimator by Forni et al. (2017) involves several estimation steps in both time and frequency domain. In this paper we propose non-parametric estimators of the common components and the common shocks directly in the frequency domain. Our model does not rely on the additional the additional assumptions in Forni et al. (2017), and our estimation method is computationally simpler and faster.

14:30 **Random matrices and spectral clustering for modeling high-dimensional self-similar systems**
Gustavo Didier

Abstract: Scale invariance (self-similarity) is a fundamental feature of the long-term behavior of many stochastic systems. The modeling of scale invariance in high dimensions is a frontier subject in statistical theory. Moreover, it impacts several fields of application, such as network traffic and neuroscience, by means of its inherent connection with large-dimensional data sets ("Big Data"). In this talk, we show that wavelet random matrices provide a natural framework for the study of scale invariance in high dimensions. In addition, we show that a spectral clustering-based algorithm can greatly improve finite-sample estimation performance. No background on the subject will be assumed.

15:00 **A semi-parametric approach for clustering high-dimensional, non-stationary, auto-correlated time series**
Qiyuan Wang, Giovanni Motta

Abstract: In this paper, we develop a novel semi-parametric estimation algorithm for accurately estimating time-varying mean and variance in autoregressive (AR) models. Utilizing B-splines with generalized least square (GLS) estimation for smooth parametrization and weighted least squares (WLS) for more precise estimation, our approach addresses the challenges posed by time-varying dynamics in time series data. The covariance matrix in our GLS estimation of the spline coefficients is iteratively updated by calculating it through the WLS estimation of the AR coefficients in a band-limited manner. Meanwhile, we propose a new autoregressive model that incorporates time-varying variance with a finite bounded envelope function and introduces a novel method to estimate it through splines. Additionally, the order of the AR model is determined through a generalized Bayesian Information Criterion (GBICp) that incorporates prior information. The effectiveness of our methodology is demonstrated through extensive simulations and applications to real-world Electro-cardiograms (ECGs) data, showcasing significant improvements in the dimension reduction while preserving major features for high accuracy clustering tasks.

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited 8

Identification and inference in semi- and non-parametric econometric models

Organizer: Karim Chalak
Chair: Karim Chalak
Room: Grande Auditorio

16:00 **Inference for Regression with Variables Generated from Unstructured Data**
Timothy Christensen

Abstract: The leading strategy for analyzing unstructured data uses two steps. First, latent variables of economic interest are estimated with an upstream information retrieval model. Second, the estimates are treated as "data" in a downstream econometric model. We establish theoretical arguments for why this two-step strategy leads to biased inference in empirically plausible settings. More constructively, we propose a one-step strategy for valid inference that uses the upstream and downstream models jointly. The one-step strategy (i) substantially reduces bias in simulations; (ii) has quantitatively important effects in a leading application using CEO time-use data; and (iii) can be readily adapted by applied researchers.

16:30 **Inference on High Dimensional Selective Labeling Models**
Shakeeb Khan, Elie Tamer, Qingsong Yao

Abstract: A class of simultaneous equation models arise in the many domains where observed binary outcomes are themselves a consequence of the existing choices of of one of the agents in the model. These models are gaining increasing interest in the computer science and machine learning literatures where they refer the potentially endogenous sample selection as the selective labels problem. Empirical settings for such models arise in fields as diverse as criminal justice, health care, and insurance. For important recent work in this area, see for example Lakkaraju et al. (2017) and Kleinberg et al. (2018), where the authors focus on judicial bail decisions, and where one observes the outcome of whether a defendant filed to return for their court appearance only if the judge in the case decides to release the defendant on bail. Identifying and estimating such models can be computationally challenging for two reasons. One is the nonconcavity of the bivariate likelihood function, and the other is the large number of covariates in each equation. Despite these challenges, in this paper we propose a novel distribution free estimation procedure that is computationally friendly in many covariates settings. The new method combines the semiparametric batched gradient descent algorithm introduced in Khan et al. (2022) with a novel sorting algorithms incorporated to control for selection bias. Asymptotic properties of the new procedure are established under increasing dimension conditions in both equations, and its finite sample properties are explored through a simulation study and an application using similar judicial bail data.

17:00 **Higher Order Moments for Differential Measurement Error, with Application to Tobin's q and Corporate Investment**
Karim Chalak, Daniel Kim

Abstract: We extend the classical measurement error model to allow the proxy for the latent vector to directly affect the response of interest, thereby violating the proxy exclusion restriction. We discuss several settings in which this type of differential measurement error occurs, and we show that higher order moments can partially identify the model parameters. In the leading case of a scalar latent variable, the identification set consists of two points. The lack of point identification occurs because relaxing the proxy exclusion restriction renders the latent variables and the proxy errors indistinguishable. However, simply signing the effects of the latent variables or distinguishing between certain moments of the latent variables and errors point identifies the model parameters. We characterize a closed form solution for the identification set in the scalar case. We propose a closed-form plug-in estimator as well as a generalized method of moments estimator. After conducting simulations, we apply our framework to estimate the firm investment equation using Tobin's q as a proxy for marginal q. We relax the standard specification to allow Tobin's q to directly affect investment, reflecting the influence of the financial market on the firms' management decisions. We find this feedback effect to be significant in both points of the identification region and more substantial when Tobin's q is a more accurate proxy for marginal q.

17:30 **Doubly Robust Bayesian Difference-in-Differences Estimators**
Christoph Breunig, Ruixuan Liu, Zhengfei Yu

Abstract: We propose a double robust Bayesian inference procedure for estimating the average treatment effect on the treated (ATT) within the difference-in-differences research design. Our robustification of the Bayesian procedure involves two important modifications: first, adjusting the prior distributions of the conditional mean function, and second, correcting the posterior distribution of the resulting ATT. We prove asymptotic equivalence between our Bayesian estimator and efficient frequentist estimators by establishing a new semiparametric Bernstein-von Mises theorem under double robustness. That is, the lack of smoothness in conditional mean functions can be compensated for by regularity of the propensity score, and vice versa. Consequently, the Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, our robust Bayesian procedure leads to a significant reduction in bias for point estimation and accurate coverage of confidence intervals, especially when the dimensionality of covariates is large relative to the sample size and the underlying functions become complex.

Advancements in semiparametric and large-scale inference
Organizer: Olga Klopp
Chair: Olga Klopp
Room: Pequeno Auditorio

16:00 **Sparse additive models with discrete optimization**
Peter Radchenko

Abstract: We will discuss recent applications of discrete optimization techniques in high-dimensional additive models. While there exist appealing approaches based on convex relaxations and nonconvex heuristics, we will focus on optimal solutions for the L0-regularized formulation, a problem that is less explored due to computational challenges. The proposed methodology covers nonparametric sparse additive modelling with smooth components and allows for pairwise interactions. Experiments based on the US Census Planning Database demonstrate that our methods automatically identify useful interactions among key factors that have been reported in earlier work by the US Census Bureau. In addition to being useful from an interpretability standpoint, our models lead to predictions that are comparable to popular black-box machine learning methods based on gradient boosting and neural networks.

16:30 **Optimal convex M-estimation via score matching**
Oliver Feng, Min Xu, Yu-Chun Kao, *Richard Samworth*

Abstract: In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. Our semiparametric approach targets the best decreasing approximation of the derivative of the log-density of the noise distribution. At the population level, this fitting process is a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. The procedure is computationally efficient, and we prove that our procedure attains the minimal asymptotic covariance among all convex M-estimators. As an example of a non-log-concave setting, for Cauchy errors, the optimal convex loss function is Huber-like, and our procedure yields an asymptotic efficiency greater than 0.87 relative to the oracle maximum likelihood estimator of the regression coefficients that uses knowledge of this error distribution; in this sense, we obtain robustness without sacrificing much efficiency. Numerical experiments confirm the practical merits of our proposal.

17:00 **Nonparametric Maximum Likelihood Estimation of Monotone Binary Regression Models under Weak Feature Impact**
Dario Kieffer, *Angelika Rohde*

Abstract: Nonparametric maximum likelihood estimation in monotone binary regression models is studied when the impact of the features on the labels is weak. We introduce a mathematical model that describes the weak feature impact. Consistency of the nonparametric maximum likelihood estimator (NPMLE) in Hellinger distance as well as its pointwise, L^1 and uniform consistency are proved in this model. Moreover, rates of consistency and limiting distribution of the NPMLE are derived. They are shown to exhibit a phase transition depending on the level of feature impact. Statistical properties of functionals in the weak feature impact scenario are also discussed.

17:30 **Dynamic Topic Model**
Cristina Butucea, Nayel Bettache, Tracy Ke

Abstract: Topic models are a widely used method to automatically produce lower dimensional latent variables explaining the output variables. It has developed merely in natural language processing (NLP) where a large number of documents can be grouped into a smaller number of topics, but it is also applied in, e.g., collaborative filtering, social networks and population genetics. It is solved by factorisation algorithms of the output matrix into the product of a word-topic matrix and a topic-document matrix under the small rank assumption. Here, by analogy to the periodic publication of documents in the media (news, research papers, etc.), we assume that an auto-regressive process drives the time evolution of the topic-document matrix, while the word-topic matrix does not change with time and verifies an anchor word assumption. We provide estimators of the parameters driving the auto-regressive process and give non-asymptotic bounds for our estimators under mild assumptions.

Structured nonparametric models
Organizer: Maria Dolores Martinez-Miranda
Chair: Maria Dolores Martinez-Miranda
Room: Sala Polivalente 1.1

16:00 **Robust and flexible model selection for local linear conditional survival function estimation**
Dimitrios Bagkavos, Jens Perch Nielsen, Montserrat Guillen

Abstract: A novel model selection approach is proposed in the context of double smoothed local linear multivariate conditional survival function estimation for right censored data. The technique is based on the intuition that inclusion of a relevant variable is expected to decrease the estimator's Mean Integrated Square Error (MISE) while, contention of an irrelevant variable results in an approximately equal or greater MISE. This, in combination with the fact that a model employing only relevant variables is expected to be asymptotically unbiased, suggests dropping a variable from the model when the MISE is greater (or the same) as the MISE of the model without it. Fair comparison of candidate models is ensured by deliberately oversmoothing irrelevant variables so as to eliminate their effect in the model's MISE. We show that, on an individual by individual basis, this approach yields the most accurate model in terms of lowest MISE. The benefits of all methodological advances are illustrated with the analysis of a motivating real-world dataset on credit risk.

16:30 **A semiparametric infinite-dimensional approach for factor analysis and dymamical multiple regression on manifolds**
María Dolores Ruiz-Medina

Abstract: This work addresses the problem of dimensionality and multicolinearity in the context of functional linear mixed models on a manifold. The presence of several functional covariates, correlated in time, affecting the evolution of the response, when uncertainly is modelled by a linear functional equation governing the dynamics of the random effect, is considered. A semiparametric framework is adopted in the estimation of the unknown dependence structure of the functional random effect on the manifold. Kernel invariance against the motions of the manifold, in the framework of Lies groups, is exploited to obtaining an universal optimal (removing multicolinearity) biorthogonal decomposition of the dynamical functional covariates and random effect, leading to an important dimension reduction. Manifold-scale dependent analysis is achieved through time in a nonparametric framework covering the case of scale-varying long memory. A simulation study is undertaken, and some real data applications are discussed to illustrate the applicability of the proposed estimation methodology.

17:00 **A Complete Framework for Model-Free Difference-in-Differences Estimation**
Stefan Sperlich

Abstract: We propose a complete framework for data-driven difference-in-differences analysis with covariates, in particular nonparametric estimation and testing. We start with simultaneously choosing confounders and a scale of the outcome along identification conditions. We estimate first heterogeneous treatment effects stratified along the covariates, then the average effect(s) for the treated. We provide the asymptotic and finite sample behavior of our estimators and tests, bootstrap procedures for their standard errors and p-values, and an automatic bandwidth choice. The pertinence of our methods is shown with a study of the impact of the Deferred Action for Childhood Arrivals program on educational outcomes for non-citizen immigrants in the US. We also discuss extensions to the inclusion of 'only moderators' and to staggered treatment.

17:30 **Smooth backfitting for additive hazard rates**
Munir Eberhardt Hiabu, Stephan Bischofberger, Enno Mammen, Jens Nielsen

Abstract: Smooth backfitting was first introduced in an additive regression setting via a direct projection alternative to the classic backfitting method in a past study. The original smooth backfitting concept is translated to a survival model considering an additively structured hazard. The model allows for censoring and truncation patterns occurring in many applications such as medical studies or actuarial science. The estimators are shown to be a projection of the data into the space of multivariate hazard functions with smooth additive components. Hence, the hazard estimator is the closest nonparametric additive fit even if the actual hazard rate is not additive. This is different to other additive structure estimators where it is not clear what is being estimated if the model is not true. The full asymptotic theory is provided for the estimators. An implementation of estimators is proposed that shows good performance in practice.

Bayesian sparse learning in high-dimensional problems
Organizer: Surya Tokdar
Chair: Surya Tokdar
Room: Sala Polivalente 1.2

16:00 **Bayesian Covariance Estimation for Multi-group Matrix-variate Data**
Elizabeth Bersson

Abstract: Multi-group covariance estimation for matrix-variate data with small within-group sample sizes is a key part of many data analysis tasks in modern applications. To obtain accurate group-specific covariance estimates, shrinkage estimation methods which shrink an unstructured, group-specific covariance either across groups towards a pooled covariance or within each group towards a Kronecker structure have been developed. However, in many applications, it is unclear which approach will result in more accurate covariance estimates. In this article, we present a hierarchical prior distribution which flexibly allows for both types of shrinkage. The prior linearly combines shrinkage across groups towards a shared pooled covariance and shrinkage within groups towards a group-specific Kronecker covariance. We illustrate the utility of the proposed prior in speech recognition and an analysis of chemical exposure data.

16:30 **Deep horseshoe Gaussian processes**
Ismaël Castillo, Thibault Randrianarisoa

Abstract: Deep Gaussian processes have recently been proposed as natural objects to fit, similarly to deep neural networks, possibly complex features present in modern data samples, such as compositional structures. Adopting a Bayesian nonparametric approach, it is natural to use deep Gaussian processes as prior distributions, and use the corresponding posterior distributions for statistical inference. We introduce the deep Horseshoe Gaussian process, a new simple prior based on deep Gaussian processes with a squared-exponential kernel, that in particular enables data-driven choices of the key lengthscale parameters. For nonparametric regression with random design, we show that the associated tempered posterior distribution recovers the unknown true regression curve optimally in terms of quadratic loss, up to a logarithmic factor, in an adaptive way. Here `adaptation' is obtained simultaneously both with respect to the smoothness of the regression function and to a possible lower-dimensional structure in terms of compositions. The dependence of the rates in terms of dimension is explicit, allowing in particular for input spaces of dimension increasing with the number of observations.

17:00 **Bayesian inference in high-dimensional mixed frequency regression**
Kshitij Khare

Abstract: Technological advancements in recent years have enabled organizations to collect, organize, store and analyze very large amounts of data from variables that are available at different temporal frequencies - eg. monthly, weekly, daily. Such data is commonly referred to as mixed frequency time series data. In the first part of the talk, we will focus on mixed frequency regression, where the response variable and the covariates are available at different frequencies (for example, quarterly vs. monthly). We will present novel Bayesian methodology for (sparse) estimation of the regression coefficients and of the (autoregressive) lag length using a Bayesian nested spike-and-slab framework. This is joint work with Satyajit Ghosh and George Michailidis.

17:30 **Bayesian Variable Selection in High-dimensional Settings with Grouped Covariates**
Minerva Mukhopadhyay

Abstract: Traditional Bayesian variable selection methods fail to yield satisfactory results in the normal linear regression setup when the number of covariates is much larger than the sample size, and the covariates form correlated groups. In such situations, sparsity exists within and between groups so that the response variable is not related to an entire group of covariates on an all or none basis. We extend Zellner's g-prior for regression parameters and the hierarchical uniform prior for models, making them appropriate for this framework, and investigate the variable selection consistency for the proposed method under fairly general conditions. In high-dimensional settings, implementation of Bayesian variable selection is extremely challenging due to the vast model space. The traditional stochastic search variable selection (SSVS) algorithms tend to get slow and inefficient under a high-dimensional correlated setup. We modify the existing SSVS algorithms using a class of group importance probabilities, termed as group-SIS (GSIS). Consequently, a novel stochastic search variable selection algorithm called group-informed variable selection algorithm (GiVSA) is proposed, which efficiently explores the model space without discarding any of the covariates in an initial screening. The performance of the proposed prior with the implementation of GiVSA is validated using a variety of numerical examples.

Statistical methods for geometric inference and set estimation
Organizer: Beatriz Pateiro-López
Chair: Beatriz Pateiro-López
Room: Sala Polivalente 1.3

16:00 **Wasserstein convergence of persistence diagrams on generic manifolds**
Vincent Divol

Abstract: Persistence diagrams (PDs) are routinely used in Topological Data Analysis to describe the topology of a sample in a multiscale fashion. They consist of a multiset of points in the upper half-plane, where each point in the PD intuitively corresponds to a topological feature of the underlying point cloud. When the sample lies on a submanifold of the Euclidean space, the PD of the sample (with respect to the Cech filtration) is known to be separated into two parts. A small number of points in the PD, which lie far away from the diagonal of the upper half-plane, correspond to the PD of the underlying manifold. On the other hand, a large collection of points lying close to the diagonal informally represents "topological noise." We provide a complete asymptotic description of the structure of this topological noise in the case where the sample lies on a generic submanifold. In particular, we offer limit laws for the total persistence of such PDs and prove convergence results with respect to Wasserstein distances. This generalizes previous results proven by Wolfgang Polonik and me in the case of points sampled in the cube [0,1]^m. This work is a joint collaboration with Charles Arnal and David Cohen-Steiner.

16:25 **Statistical difficulty of support estimation and dimensionality reduction**
Clément Levrard, Eddie Aamari, Catherine Aaron, Clément Berenfeld

Abstract: A common assumption in modern nonsupervised ML is the so-called 'Manifold assumption', that roughly supposes that data points, though observerd in a high-dimensional ambient space, are in fact elements of a low-dimensional hidden structure (a manifold). In this setting, we focus on two specific tasks: retrieving the hidden structure (support estimation) and embedding data points in a low-dimensional Euclidean space (dimensionality reduction). Combining results from a recent line of work, in collaboration with E. Aamari, C. Aaron and C. Berenfeld, I will briefly expose how these two tasks are connected, and try to characterize their statistical difficulty (minimax rates on appropriate models) to partially answer the question: is dimensionality reduction easier than support estimation?

16:50 **Confidence Regions for Filamentary Structures**
Wanli Qiao

Abstract: Filamentary structures, also called ridges, generalize the concept of modes of density functions and provide low-dimensional representations of point clouds. Using kernel type plug-in estimators, we give asymptotic confidence regions for filamentary structures based on two bootstrap approaches: multiplier bootstrap and empirical bootstrap. Our theoretical framework respects the topological structure of ridges by allowing the possible existence of intersections. Different asymptotic behaviors of the estimators are analyzed depending on how flat the ridges are, and our confidence regions are shown to be asymptotically valid in different scenarios in a unified form.

17:15 **Highest density region estimation for manifold data**
Diego Bolón, Rosa M. Crujeiras, Alberto Rodríguez-Casal

Abstract: Highest density regions (HDRs) are defined as the sets where the density function of the data exceeds a fixed (and usually high) level. Given a data sample, estimating the HDRs of the underlying density is a useful tool for data modelling, exploration, and visualization. Some of the multiple applications of this technique are the localization of minefields based on aerial observations, the analysis of seismic data, and the detection of outliers within a sample. Estimating HDRs for Euclidean data (uni or multidimensional) has been widely considered in the statistical literature. However, HDRs estimation in other domains has received very little attention in comparison, with some recent proposals of HDR estimators for data on compact manifolds. The most studied approach for estimating HDRs on manifolds is a plug-in method, that is, the HDR estimator is directly defined as the HDR of an estimator of the underlying density. Although this estimation technique is conceptually simple, it ignores the geometrical structure of the problem. If one knows that the HDRs fulfill some geometric property (e.g. some shape condition), there is no guarantee that the plug-in estimator will satisfy it too. Trying to overcome these issues, a new non-parametric HDR estimator for manifold data is introduced. The new approach combines an estimator of the underlying density with some a priori geometric information, which is appropriately included in the estimation method to simplify the final form of the estimator.

17:40 **Two sample testing for isometry of two manifolds**
Wolfgang Polonik, Eunseong Bae

Abstract: We are presenting a two-sample test for isometry of two manifolds by combing ideas from spectral analysis and topological data analysis. More specifically, the proposed test statistic is the bottleneck distance between the persistence diagrams generated by estimated heat kernel signatures (HKS) of the two manifolds. Under certain assumptions, the asymptotic normality of an appropriately normalized version of this test statistic can be derived. A bootstrap approach is then being used to conduct the test. One of the technical ingredients into the proofs is a uniform convergence rate of a kernel density estimator (KDE) on a closed manifold, which might be of independent interest. We will discuss and motivate the basic approach to the hypothesis testing problem and also discuss some challenges, in particular related to the bias of the KDE caused by the curvature of the manifold. This is joint work with Eunseong Bae.

Advances in functional data analysis
Organizer: Michelle Carey
Chair: Michelle Carey
Room: Sala Polivalente 1.4

16:00 **Regression in quotient metric spaces with a focus on elastic curves**
Lisa Steyer, Almond Stöcker, *Sonja Greven*

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

16:30 **Space-time regression with non-stationary PDE penalization for the analysis of mobile phone data**
Eleonora Arnone, Mara Sabina Bernardi, Laura Maria Sangalli, Piercesare Secchi

Abstract: We analyze mobile phone data collected in the metropolitan area of Milan, combining the phone data with information about the road networks of the Lombardia region. We use spatio-temporal regression with Partial Differential Equation (PDE) penalization. This technique allows the inclusion of specific information on the phenomenon under study through the definition of a non-stationary anisotropy modeled via a PDE, defined over the spatial domain of interest. The non-stationary parameters of the PDE are estimated starting from the road network of the area, allowing a preferred direction of smoothing along the highways, and an isotropic smoothing far from the main roads, where no information on preferred movements is available.

17:00 **Statistical Analysis of Collections of Networks**
Catherine Higgins, Michelle Carey, Hulin Wu

Abstract: Networks serve as powerful tools to capture and represent information from complex, high-dimensional data. The statistical analysis of collections of networks, where the network itself is treated as the fundamental unit of observation, is important across a variety of domains such as gene regulatory networks, social networks, brain activity networks, financial networks, and transport networks. However, extending classical statistical methods to network-based data poses challenges due to the non-Euclidean nature of networks, defined by vertices and edges. A key issue in network analysis is quantifying similarity or distance between networks of varying sizes and types such as directed/undirected, weighted/unweighted or when node correspondences are unknown. To address this need, we propose a novel method utilizing topological data analysis, specifically persistent homology, to characterize and statistically analyze network-based datasets. Persistent homology offers a method to compute topological features persisting across multiple scales, providing insights into the structure of networks. As a result, fundamental statistical concepts can be achieved for collections of networks, which we demonstrate on simulated networks arising from different connectivity structures. We also apply the proposed approach to a sample of gene regulatory networks from two distinct groups and devise a hypothesis test to discern differences among network groups.

17:30 **Block testing in precision matrix for functional data analysis**
Alessia Pini, Marie Morvan, Madison Giacofci, Valerie Monbet

Abstract: We propose a method to test conditional linear independence between portions of the domain of functional data. In particular, we assume that data can be described by means of a (possibly high-dimensional) B-splines basis expansion, such that coefficients of the basis expansion are directly related to the parts of the domain where the support of basis functions is strictly positive. We further assume that the domain can be partitioned into regions of interest. This is usually the case when it is possible to identify landmarks on functional data (regions of interests are intervals whose endpoints are landmarks), or when some information is available on the domain. In such a case, we expect the precision matrix to have a block structure, where blocks correspond to elements of the partition. So, to infer about which areas of the domain - that are related to components of the partition - are conditionally linearly independent between each other, we propose to test blocks of the precision matrix of basis coefficients. The tests are based on permutation procedure that tests if suitable blocks of the precision matrix of basis expansion coefficients are equal to zero. We introduce a suitable strategy to deal with the multiple testing issue in this setting, controlling the hereby defined block-wise error rate. We show that the procedure is able to identify the true structure of conditional dependence on simulated data and on a real case study involving tractographic data related to the infrared emission spectra of fruit purees.

Recent advances in non-and semiparametric models in survival analysis

Organizer: Ingrid Van Keilegom
Chair: Ingrid Van Keilegom
Room: Sala Polivalente 1.5

16:00 **Survival Estimation with Time-Varying Covariates Using Neural Networks**
Bingqing Hu, *Bin Nan*

Abstract: Most work in neural networks focuses on estimating the conditional mean of a continuous response variable given a set of covariates. In this talk, we consider estimating the conditional distribution function using neural networks for censored survival data. The algorithm is built upon the data structure particularly constructed for the Cox regression with time-dependent covariates. Without imposing any model assumption, we consider a loss function that is based on the full likelihood where the conditional hazard function is the only unknown parameter, for which unconstraint optimization methods can be applied. Through simulation studies, we show the proposed method possesses desirable performance, whereas the partial likelihood method yields biased estimates when model assumptions are violated.

16:30 **Cumulative Incidence Function Estimation Using Population-Based Biobank Data**
Malka Gorfine, David M. Zucker, Shoval Shoham

Abstract: Many countries have established population-based biobanks, which are being used increasingly in epidemiological and clinical research. These biobanks offer opportunities for large-scale studies addressing questions beyond the scope of traditional clinical trials or cohort studies. However, using biobank data poses new challenges. Typically, biobank data is collected from a study cohort recruited over a defined calendar period, with subjects entering the study at various ages falling between c_L and c_U. This work focuses on biobank data with individuals reporting disease-onset age upon recruitment, termed prevalent data, along with individuals initially recruited as healthy, and their disease onset observed during the follow-up period. We propose a novel cumulative incidence function (CIF) estimator that efficiently incorporates prevalent cases, in contrast to existing methods, providing two advantages: (1) increased efficiency, and (2) CIF estimation for ages before the lower limit, c_L.

17:00 **A competing risks analysis with cause-specific cure**

Eni Musta, Tijn Jacobs, Marta Fiocco

Abstract: We consider survival analysis in the presence of competing events. Standard methods to deal with competing risks assume that all subjects are susceptible to all events and only few recent papers try to accommodate the possibility of being immune ("cured") to a subset of the risks or all of them simultaneously. We propose a general model for two competing events and cause-specific cure for each event. Then, the previously mentioned settings considered in the literature (cure for only one event or for both simultaneously) become particular cases of this more general model. Our research is motivated by the question: can we identify the relation between the two cause-specific cure statuses without making apriori restrictions? We consider a logistic model for the cure probabilities and a semiparametric Cox model for the cause-specific hazards. First we discuss which quantities can be identified from the data and under what assumptions. In addition, we propose an estimation procedure based on the EM algorithm and investigate both asymptotic and finite sample performance of the method. The method is illustrated through an application to consumer loans data for which the competing events are default and prepayment.

17:30 **Conditional C-index for survival data with a cure fraction**

Bo Han, Ingrid Van Keilegom, *Juan Carlos Pardo-Fernandez*

Abstract: When analyzing survival data, it often occurs that some individuals never experience the event of interest. These individuals are called cured and cure models are then used to take into account this situation. In particular, in this talk we will consider a mixture cure model, which combines the probability of being uncured (also called incidence) and the survival function of the uncured patients (also called latency). In practice, risk scoring systems of latency and incidence are crucial elements for identifying relevant biomarkers and treatment strategies. Concordance measures that discriminate higher-risk subjects from lower-risk subjects are valuable tools to evaluate the overall performance of risk scoring systems. In contrast to conventional concordance measures, conditional concordance measures are proposed in this talk to provide comprehensive assessment of fitted cure models for particular values of a set of covariates. Specifically, we will consider the conditional version of the concordance index or C-index to evaluate the discrimination capacity of risk factors for both the latency and the incidence. Non- and semi-parametric modelling strategies are proposed to estimate and perform inferences about the conditional C-index. Simulation studies demonstrate that our proposal has a promising performance in finite samples. An application to real data is presented for illustrating the methodology.

18:00 - 19:00 Posters

Scalar-on-Shape Regression Models in Functional Data Analysis

Sayan Bhadra, Anuj Srivastava

Abstract: Our work studies shape-based functional data analysis (FDA), a branch of FDA where the shapes of functions are considered their preeminent feature. Specifically, it focuses a family of scalar-on-shape regression models that consider only the shapes of predictor functions and discard their phases as nuisance. This is different from the traditional scalar-on-function regression models where the full predictor functions are used. We develops shape-regression models using an elastic Riemannian metric to impose invariant inner-products and nonlinear index functions to capture complex predictor-response relationships. This formulation also leads to a novel definition of the regression phase of functional data; the regression phase is defined as times warpings of predictor functions that optimizes their predictive power in a regression model. We also analyze estimation of regression parameters and conclude that the traditional optimization approach leads to biased estimators. We demonstrate this framework using a number of simulated and real-data examples, including the prediction of COVID outcomes using daily infection-rate curves as predictors.

Modal Regression with Missing Response Data

Tomás R. Cotos-Yáñez, Rosa M. Crujeiras, Ana Pérez-González

Abstract: Modal regression estimates local modes of the conditional distribution of a response variable Y conditional on X=x. It is an alternative method to the classic mean regression that has gained much importance in recent decades due to its suitability, for example, when the conditional distribution has heavy tails, is not symmetrical or has more than one mode. On the other hand, we often encounter incomplete samples where the study of the behavior of any estimator under missing information is crucial to make decisions about the different missing data methodologies that can be applied. In this work, we adapt some methodologies used in modal regression to the context of missing response data: using just complete observations incorporating weitghts to the estimator, or imputing the missing response (simple or multiple imputation). The performance of the different estimated estimators is analysed in an extensive simulation study and an application to real data is also included.

Goodness-of-fit tests for circular data based on a Parzen-Rosenblatt type estimator

Carlos Tenreiro

Abstract: Given an independent and identically distributed sample of angles from some absolutely continuous circular random variable with unknown probability density function f, in this work we study the problem of testing the hypothesis on whether f belongs to a given parametric class of densities. For this purpose we consider a Bickel-Rosenblatt type test statistic (L^2 distance) based on the Parzen-Rosenblatt type estimator for circular data recently introduced and studied in Tenreiro (2022, doi: 10.1080/10485252.2022.2057974; 2023, doi: 10.1080/03610926.2023.2264996). The asymptotic behaviour of the proposed test procedure for fixed and non-fixed bandwidths is studied. From a finite sample point of view the power performance of the tests associated with different bandwidths depends on the considered bandwidth which acts as a tuning parameter. The automatic selection of this tuning parameter, which choice is crucial to obtain a performing test procedure, is also addressed in this work.

Estimating asymptotic independence on the lower tail

Marta Ferreira

Abstract: Extreme values theory has been growing in its application to various areas, especially in the analysis and modeling of phenomena involving risk. The assessment of the dependence between losses on financial market indices is one of the application examples. In this work we address the estimation of a measure of residual dependence (or asymptotic independence) and compare different estimators through a simulation study. An illustration with financial data is also presented.

Multivariate asymptotic test of pairwise independence for orientations with the same symmetry group

Iva Karafiátová, Jakub Staněk, Zbyněk Pawlas

Abstract: The subject of this work is the independence of r-tuples of orientations of symmetrical objects in three-dimensional space. Beyond its applications in crystallography and material science, such objects occur in various fields. The characteristics of random orientations and the properties of their estimators are discussed. The main contribution is constructing multivariate asymptotic tests of pairwise independence based on the theory of U-statistics and a novel definition of covariance between two orientations. The finite-sample performance of the tests is assessed through a simulation study, investigating their power on three models for r-tuples of orientations and comparing them to a permutation test. The application of the test is demonstrated on a dataset of polycrystalline material with cubic symmetry of the crystal lattice.

Inference on Data with Both Multiplicative and Additive Measurement Error

Yuxiang Zong, Yinfu Liu, Yanyuan Ma, Ingrid Van Keilegom

Abstract: Measurement errors are omnipresent in many fields and can lead to serious problems in statistical analysis. In the literature, measurement errors are often assumed to be either additive or multiplicative. We consider the case where a variable is subject to both additive and multiplicative errors. We establish the identifiability and propose a moment-based estimator for the variances of these two types of errors, which is shown to be consistent. We further derive the asymptotic distribution of the estimator and conduct hypothesis tests to examine the existence of the two types of errors. A likelihood-based method is developed to approximate the density of the error-prone variable. We apply our strategy in the context of linear regression and study its effect on the estimation of regression parameters in combination with Regression Calibration and Simulation Extrapolation. The proposed methodology is numerically investigated through simulations and is implemented in a real data application.

Semiparametric Generative Invariance

Carlos García Meixide, David Ríos Insua

Abstract: This work enhances the current statistical comprehension of causality by demonstrating that domain generalization is achievable without heterogeneous data sources and without specific assumptions regarding the strength of perturbations or support overlap. We present a new estimator for predicting outcomes in different distributional settings under hidden confounding without relying on instruments or exogenous variables. The population definition of our estimator identifies causal parameters belonging to certain complexity classes, whose empiricals lead to a generative model capable of replicating the true probability law of the outcome given the covariate distribution at test stage beyond (suboptimal) do-interventional conditionals. We show that the probabilistic alignment between our proposed method and true test distributions is uniformly the best across various interventions.

Nonparametric methods for the extremal index estimation

Dora Prata Gomes, Manuela Neves

Abstract: Extreme value theory offers limiting distributions for rare events across a broad range of stationary time series. This theory deals not only with the magnitude of extremes but also with the frequency of their occurrence. Our attention in this context is directed towards understanding the clustering pattern of extremes. Frequently, it is observed that extremes, such as temperatures, water levels, wind speeds, or financial time series, exhibit clustering behavior over time. In other words, they do not occur randomly, as one would expect from a Poisson process. Extreme value theory can incorporate such clustering tendencies through the extremal index, while maintaining the unchanged shape of the Generalized Extreme Value (GEV) distribution. Therefore, the estimation of the extremal index is a topic of great interest. Its estimation has been addressed by numerous authors; however, several challenges persist. One such challenge involves determining the appropriate number of upper order statistics to consider in semiparametric estimation. Overall, the concept of employing nonparametric methods for estimating the extremal index remains relatively underexplored, promising to offer novel insights. We will present several results from a simulation study, as well as an application to hydrological data. This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (Center for Mathematics and Applications) and UIDP/00006/2020 (https://doi.org/10.54499/UIDP/00006/2020) (CEAUL- Centre of Statistics and its Applications).

Bayesian Additive Regression Trees in Complex Survey

Abhishek Mandal

Abstract: Complex surveys have garnered substantial significance across diverse domains, spanning social sciences, public health, and market research. Their pivotal role lies in furnishing representative estimations while adeptly addressing the intricacies of survey design effects. When faced with the intricate complexities arising from the unknown effects of various covariates, parametric approaches may prove insufficient in handling the nuances associated with survey design impacts. Additionally, the Gaussian error distributional assumption would be inappropriate in many applications where the response distribution is heavy-tailed or skewed. This paper introduces the Bayesian Additive Regression Trees (BART) framework—a potent and adaptable approach tailored for analysing intricate survey data, specifically with subject weights. We propose an extension of BART to model heavy-tailed and skewed error distribution while considering subject weights. Its ability to account for the survey design features, handle non-linearity, and provide uncertainty estimates makes it a valuable tool for researchers and practitioners working with complex survey data. We illustrate the advantages of our methods through a simulation study and analysis of dental data from the NHANES study.

Modeling multivariate spatial dependency with copulas: a novel approach

Manuel Úbeda-Flores

Abstract: A new approach to modeling multivariate spatial dependency using copulas ---bivariate distribution functions with uniform margins in [0,1]--- has gained popularity in spatial statistics. In spatial modeling, copulas can be employed to describe the nonlinear spatial dependence between pairs of points associated with a given univariate spatial random field. This is particularly useful when close point pairs exhibit a dependence structure close to perfect, while independence is observed for pairs separated by large distances. The spatial dependence can be characterized by a distant-dependent bivariate copula, and a spatial copula corresponding to each distance class can be constructed as a convex combination of bivariate copulas. We introduce a novel mixture copula framework designed to effectively model dependencies among multiple spatially correlated variables.

Local logistic regression for dimension reduction in binary classification

Touqeer Ahmad, François Portier, Gilles Stupfler

Abstract: The framework of sufficient dimension reduction has progressed steadily. However, its ability to enhance the general framework for binary classification has not received much attention yet, especially for high-dimensional data. This study addresses the dimension reduction challenge in binary classification when a large set of covariates exist. The primary objective is to define and estimate a projection onto a lower-dimensional subspace of the covariate space, tailored explicitly for checking instances of binary response. To achieve this goal, we introduce a novel dimension-reduction approach based on a localized logistic model with a penalty parameter ($\lambda = 0$ and $\lambda > 0$). We use local neighborhoods of covariate data points along with corresponding target variable instances. Our proposed approach demonstrates superiority in both synthetic data and real data applications compared to existing competitors.

Approximate Bayesian computation for Arrhenius relationship accelerated life tests

Lizanne Raubenheimer, *Neill Smit*

Abstract: Accelerated life testing can be used to estimate the life characteristics of high-reliability products, especially where conventional reliability estimation is not possible due to time and cost constraints. In an accelerated life test, products are exposed to more severe than their normal operating conditions, by applying stressors, to induce early failures. A time transformation function, which is a functional relationship between the parameters of the life distribution and the accelerated stressors, can then be used to estimate the life characteristics of the products under their normal operating conditions. The resulting models are often complicated and classical parameter estimation is not always possible. In this paper, we consider accelerated life testing for some widely used life distributions and the Arrhenius relationship as the time transformation function. A likelihood-free method, using approximate Bayesian computation, is investigated for parameter estimation and model selection. The approximate Bayesian computation method is compared to classical methods, such as maximum likelihood estimation, in a simulation study.

Explainable Deep Learning: a methodology to train Generalized Additive Model with deep neural networks

Ines Ortega-Fernandez, Marta Sestelo

Abstract: Neural networks have become increasingly popular due to their remarkable performance across various domains, including computer vision, anomaly detection, and cybersecurity. However, the inherent black-box nature of deep neural network architectures poses challenges in understanding their decision-making processes. In this work, we present a neural network topology inspired by generalized additive models (GAM), which trains independent neural networks to estimate the effect of each covariate on the response variable, leading to the creation of a highly accurate and interpretable deep learning-based Generalized Additive Neural Network (GANN) model. We leverage the backfitting and local scoring algorithms to train a highly accurate yet interpretable deep learning model, showcasing the effectiveness of the method in a real Denial of Service cyberattack detection scenario.

Functional Data Analysis for Predicting Landed Fish Abundance per unit effort (LPUE)

Manuel Oviedo-de la Fuente, Raquel Menezes, Alexandra A. Silva

Abstract: Predicting the abundance of landed fish per unit effort (LPUE) is a critical challenge in competitive fish markets. Previous research has addressed the challenge of modelling species distribution in fisheries using various statistical methods, including time series analysis (e.g., ARIMA models), model-based geostatistics (e.g., SPDE approach, GRFs and kriging), and regression models (e.g., GLMs) to model parametrically temporal, spatial and other complex structures This study proposes an approach based on functional data analysis (FDA). FDA is a branch of statistics that focuses on the analysis of data consisting of curves or anything else that varies along a continuum. This paper addresses the challenge of variable selection by using distance correlation to investigate the relationships between predictors (functional and scalar) and the scalar response (LPUE) The proposed functional approach, specifically a functional non-parametric model (FNPM), has demonstrated promising results when applied to a real dataset (LPUE of juvenile sardine, along the northern Portuguese coast) using functional predictors such as chlorophyll-a concentration, ocean current intensity, sea surface temperature, wind speed, and wind direction and other relevant information such as sale prices at landing and calendar variables in the variable selection process These results provide decision-makers with a valuable tool to advance marine sustainability and conservation efforts by improving our understanding of the factors influencing LPUE. Acknowledgement: This research/work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01), all of them through the ERDF. Authors also acknowledge the FCT Foundation for funding their research through projects with DOI 10.54499/UIDP/00013/2020 and DOI 10.54499/UIDB/00013/2020 and MAR2030 for funding the SARDINHA2030 project.

Extrinsic Principal Component Analysis

Ka Chun Wong, Vic Patrangenaru

Abstract: We aim to develop a sustainable methodology for principal component analysis on manifolds. Instead of estimating intrinsic principal components on an object space with a Riemannian manifold structure, we propose a faster method that is working via an embedding of the object space in a numerical space, and having the resulting chord distance. This method helps us analyzing infinite dimensional planar shapes, from a new perspective. We define the extrinsic principal sub-manifolds of a random object on a manifold embedded in an Euclidian space, and their sample counterparts. The principal submanifolds are necessary for dimension data reduction of high dimensional objects beyond shapes of planar contour data, in 3D for projective shapes or 3D Kendall shapes. For applications, we retain a reasonably small number of such extrinsic principal submanifolds for a contour data sample, extracted from imaging planar data.

Clustering cumulative incidence functions with clustcurv R package.

Marta Sestelo, Luís Meira-Machado, Nora Villanueva, Javier Roca-Pardiñas

Abstract: The package clustcurv is freely available at Comprehensive R Archive Network (CRAN). At the moment, it contains functions for determining groups in multiple survival curves and regression curves. However, these functions cannot be applied to competing risk survival data. To solve this issue, new R functions have been implemented and included in a recent version of the package (available at GitHub), containing new methods for clustering cumulative incidents functions. The repository can be found at https://github.com/noramvillanueva/clustcurv.

Testing the fit of discrete response models with covariates

Leonard Santana, Simos Meintanis, Marius Smuts, Joseph Ngatchou Wandjii

Abstract: We propose goodness–of–fit tests for models of count responses with covariates. Our main focus is on the null hypothesis that the observed data come from a Poisson, a negative binomial, or a binomial regression model, but the method is fairly general allowing for the responses to follow, conditionally on covariates, any given discrete distribution. The test criteria are formulated by using the probability generating function and they are convenient from the computational point of view. Asymptotic as well as Monte Carlo results are presented. Applications on real data are also reported.

New classes of tests for the Weibull distribution using Stein's method in the presence of random right censoring

Elzanie Bothma, James Allison, Jaco Visagie

Abstract: We develop two new classes of tests for the Weibull distribution based on Stein's method. The proposed tests are applied in the full sample case as well as in the presence of random right censoring. We investigate the finite sample performance of the new tests using a comprehensive Monte Carlo study. In both the absence and presence of censoring, it is found that the newly proposed classes of tests outperform competing tests against the majority of the distributions considered. In the cases where censoring is present we consider various censoring distributions. Some remarks on the asymptotic properties of the proposed tests are included. We present another result of independent interest; a test initially proposed for use with full samples is amended to allow for testing for the Weibull distribution in the presence of censoring. The techniques developed in the paper are illustrated using two practical examples.

On a new class of tests for the Pareto distribution using Fourier methods

Lethani Ndwandwe, James Allison, Marius Smuts, Jaco Visagie

Abstract: We propose new classes of tests for the Pareto type I distribution using the empirical characteristic function. These tests are U and V statistics based on a characterization of the Pareto distribution involving the distribution of the sample minimum. In addition to deriving simple computational forms for the proposed test statistics, we prove consistency against a wide range of fixed alternatives. A Monte Carlo study is included in which the newly proposed tests are shown to produce high powers. These powers include results relating to fixed alternatives as well as local powers against mixture distributions. The use of the proposed tests is illustrated using an observed data set.

On classes of consistent tests for the Pareto distribution based on a characterization involving order statistics

Thobeka Nombebe, James Allison, Joseph Ngatchou--Wandji, Leonard Santana

Abstract: We propose new classes of goodness-of-fit tests for the Pareto Type I distribution. These tests are based on a characterization of the Pareto distribution involving order statistics. We derive the limiting null distribution of the tests and also show that the tests are consistent against fixed alternatives. The finite-sample performance of the newly proposed tests is evaluated and compared to some of the existing tests, where it is found that the new tests are competitive in terms of powers. The talk concludes with an application to a real-world data set, namely the earnings of the 22 highest-paid participants in the inaugural season of LIV golf.

Saturday, 29 Jun

9:00 - 11:00 Invited 9

Statistics for a wise use of machine learning

Organizer: Stefan Sperlich
Chair: Stefan Sperlich
Room: Grande Auditorio

9:00 **Fairness in Machine Learning : how AI creates and amplifies bias in the data.**
Jean-Michel Loubes

Abstract: Fairness has been at the heart of many research in Machine Learning over the recent years. In many study, an algorithm built using machine learning is said to learn, reproduce and amplify bias present in the data. We considering measures of bias (direct or indirect, i.e statistical parity measures or equal opportunity measures), the level of bias in the output of the algorithm is often worse than the initial level of bias in the data, leading to possible increase of discrimination in the algorithmic decisions. This idea has been developed in most of the research papers with many empirical proofs for the so-called {bias amplification} phenomenon using simulated or even real life data. Here we propose to provide a theoretical proofs of this bias amplification and to obtain mathematical guarantees to understand the origin of bias in machine learning algorithmic decisions.

9:30 **Statistical methods for high-throughput experimental data**
Tatyana Krivobokova, Gianluca Finocchio, Boris Maryasin

Abstract: Exploiting reactions from high-throughput experimentation using machine learning techniques is becoming state-of-the-art in organic chemistry. Many valuable data sets are being generated in order to learn about reaction conditions that are crucial for the outcomes of chemical reactions such as yields or selectivities. However, it is typically ignored, that the data from designed experiments inherit a very specific structure that needs to be taken into account in the analysis in order to make appropriate conclusions. On the example of the data from Buchwald-Hartwig Amination, which were collected for nearly 4000 reaction conditions, we demonstrate the shortcomings of used machine learning techniques and suggest a statistically rigorous approach to the analysis of such data from high-throughput experiments.

10:00 **Trends in Statistical Deep Learning**
Johannes Lederer

Abstract: Deep learning and artificial intelligence are currently among the hottest topics in science. The corresponding agenda is set by computer science and engineering, sidelining statistics almost entirely. Therefore, this talk will advocate statistical viewpoints, and it will highlight progress at the many intersections of mathematical statistics, computer science, and applications.

10:30 **The implicit bias phenomenon in deep learning**
Holger Rauhut

Abstract: Deep neural networks are usually trained by minimizing a non-convex loss functional via (stochastic) gradient descent methods. Unfortunately, the convergence properties are not very well-understood. Moreover, a puzzling empirical observation is that learning neural networks with a number of parameters exceeding the number of training examples often leads to zero loss, i.e., the network exactly interpolates the data. Nevertheless, it generalizes very well to unseen data, which is in stark contrast to intuition from classical statistics which would predict a scenario of overfitting. A current working hypothesis is that the chosen optimization algorithm has a significant influence on the selection of the learned network. In fact, in this overparameterized context there are many global minimizers so that the optimization method induces an implicit bias on the computed solution. It seems that gradient descent methods and their stochastic variants favor networks of low complexity (in a suitable sense to be understood), and, hence, appear to be very well suited for large classes of real data. Initial attempts in understanding the implicit bias phenomen considers the simplified setting of linear networks, i.e., (deep) factorizations of matrices. This has revealed a surprising relation to the field of low rank matrix recovery (a variant of compressive sensing) in the sense that gradient descent favors low rank matrices in certain situations. Moreover, restricting further to diagonal matrices, or equivalently factorizing the entries of a vector to be recovered, shows connection to compressive sensing and l1-minimization. Despite such initial theoretical results on simplified scenarios, the understanding of the implicit bias phenomenon in deep learning is widely open.

Nonparametric smoothing and regression for correlated observations

Organizers: Didier A. Girard and Sana Louhichi
Chair: Sana Louhichi
Room: Pequeno Auditorio

9:00 **Inference on volatility estimation with missing data: a functional data approach**
Mohamed Chaouch, Abdelbasset Djeniah, Amina Angelika Bouchentouf

Abstract: In the last 15 years, the capital markets have seen significant development, introducing high-frequency trading and a shift of market towards high-frequency and algorithm trading. It was always believed that high-frequency trading and automated trading were source price shocks and rising of volatility. Moreover, financial firms, that trade assets on high-frequency time scale, are not just interested in short-term forecasting of future values of financial assets, but also in measuring the uncertainty associated to such predictions through the volatility component. This paper aims to investigate nonparametric estimation of the volatility component in a heteroscedastic scalar-on-function regression model when the underlying discrete-time process is ergodic and the response variable is affected by a missing at random mechanism. First, we introduce a simplified estimator of the regression and volatility operators based on observed data only. We study their asymptotic properties, such as almost sure uniform consistency rate and asymptotic distribution. Then, the simplified estimators are used to impute the missing data in the original process in order to improve the estimation of the regression and volatility components. The asymptotic properties of the imputed estimators are also investigated. A numerical comparison between the estimators is discussed through simulated data. Finally, a real-data analysis is conducted to model the volatility of daily Brent crude oil returns using intraday, 1-minute frequency, natural gas returns.

9:30 **Non-parametric statistic to test the equality of the health concentration curve and the 45 degree line**
Taoufik Bouezmarni, Abderrahim Taamouti, Mohamed Doukali, Meryem Taleb Bendiab

Abstract: The aim of this paper is to develop a non-parametric statistic to test the equality of the health concentration curve and the 45 degree line. The test statistic is based on the comparison of the health concentration curve, C(p), and p in (0,1), using a L_2 metric. We use the non-parametric estimator for C proposed by Wastgaff (1989). We establish the asymptotic size and power properties of the test statistic. A Monte Carlo simulation study shows that our tests have good finite sample size and power properties for a variety of data generating processes and different sample sizes. Finally, we provide an empirical application to illustrate the usefulness of the proposed test.

10:00 **On kernel density estimation for dependent data on Riemannian manifolds without boundary**
Anne Francoise Yao, Vincent Monsan, Djack Guy-Aude Kouadio

Abstract: Let, X, be a variable with unknown density, f. The problem of kernel density estimation estimator (KDE) has been widely studied when X lies in an Euclidean space of dimension d (even in functional data space). However, in many situations, X is with values on a d-1 Riemannian submanifold, M (as illustrated for example in Mardia et al. (2008) and Pennec (2006)) so that any statistical information on X is linked to an (unusual) geometry of M. Concerning the problem of estimation of f, several kernel estimators has been proposed in geometry and statistics. This talk focuses on the KDE version of Pelletier (2005) which was firstly built for independent identically distributed observations. To our knowledge, the case of KDE based on dependent data has not been treated in the literature (at least theoretically). Namely, we address the KDE issue based on observations generated from a strong-mixing stationary continuous processes. We will give some asymptotic results on our estimator.

10:30 **Hyperparameters selection problems in nonparametric trend estimation: from statistics to machine learning**
Sana Louhichi

Abstract: The problem of curve estimation, such as the estimation of the regression function, appears in many applied fields for instance in pharmacokinetics, in medicine or in environmental science. In this talk, we are interested in the problem of smoothing parameter selection in nonparametric curve estimation under dependent errors. We focus on kernel estimation and the case when the errors form a general stationary sequence of martingale difference random variables where neither linearity assumption nor "all moments are finite" are required. We compare the behaviors of the smoothing bandwidths obtained by minimizing three criteria: the average squared error, the mean average squared error and a Mallows-type criterion adapted to the dependent case. We prove that these three minimizers are first-order equivalent in probability. We give also a normal asymptotic behavior of the gap between the minimizer of the average square error and that of the Mallows-type criterion. Finally, we apply our theoretical results to a specific case of martingale difference sequence, namely to some stochastic volatility sequences. Some Monte-Carlo simulations studies are conducted. The talk is based on common works with Karim Benhenni and Didier Girard.

Advances in directional statistics
Organizer: Thomas Verdebout
Chair: Rosa M. Crujeiras
Room: Sala Polivalente 1.1

9:00 **On dependence analysis for circular data**
Rosa M. Crujeiras

Abstract: We will revise in this talk some different approaches to model dependence involving circular data. Classical correlation coefficients, copula models and testing procedures, will be illustrated with simulated examples. In some settings, the existing approaches do not suceed in noticing the dependence existing in the data, that could be formed by circular-circular or circular-linear paris. We will introduce the use of correlation-distance based tests in this settings and compare their performance with the previous proposals.

9:30 **Kernel density estimation on the polysphere**
Eduardo García-Portugués, Andrea Meilán-Vila

Abstract: A kernel density estimator for data on the polysphere S^{d_1} x ··· x S^{d_r}, with r, d_1, ..., d_r ≥ 1, is introduced. Among the asymptotic properties of the estimator, particular focus is placed on the kernel theory of the estimator, which goes beyond product and von Mises–Fisher kernels. As a spin-off, a nonparametric k-sample test based on the Jensen–Shannon divergence is introduced. Numerical experiments illustrate the kernel theory and the performance of the k-sample test. The methodology is applied to the analysis of the morphology of a sample of hippocampi of infants.

10:00 **Conditional density estimation for spherical data**
María Alonso-Pena, Paula Saavedra-Nieves

Abstract: Kernel density estimation has been widely employed to estimate density functions supported on the unit (hyper)sphere. In this talk, we will consider the nonparametric estimation of conditional densities, where both the variable of interest and the conditioning variable are supported on the unit (hyper)sphere. We will present the kernel estimator of the conditional density and show some of its asymptotic properties. In addition, we will discuss the problem of selecting a data-driven smoothing parameters, necessary to obtain the estimator in practice. Simulated data will be employed to show the finite sample properties of the proposed estimator, considering different simulated models. We will also show an application of the new methodology with real data, regarding an animal escapology problem. Lastly, we will discuss some applications of the kernel estimator of the spherical conditional density to other statistical problems and further challenges.

10:30 **A novel data-based smoothing parameter for circular kernel density estimation**
Jose Ameijeiras-Alonso

Abstract: In this talk, we will introduce an innovative approach to circular kernel density estimation by proposing a novel data-based smoothing parameter. Drawing inspiration from the well-known Sheather and Jones bandwidths, we adapt them to the circular context, substituting unknown parameters with estimations derived from plug-in techniques. Theoretical foundations supporting our method are outlined, followed by an overview of the simulation study assessing the performance of our proposed selectors against established data-based smoothing parameters.

Network models and optimal prediction

Organizer: Moulinath Banerjee
Chair: Moulinath Banerjee
Room: Sala Polivalente 1.2

9:00 **UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation**
Debarghya Mukherjee, Jiawei Ge, Jiaqing Fan

Abstract: Uncertainty quantification for prediction is an intriguing problem with significant applications in various fields, such as biomedical science, economic studies, and weather forecasts. Numerous methods are available for constructing prediction intervals, such as quantile regression and conformal predictions, among others. Nevertheless, model misspecification (especially in high-dimension) or sub-optimal constructions can frequently result in biased or unnecessarily-wide prediction intervals. In this paper, we propose a novel and widely applicable technique for aggregating multiple prediction intervals to minimize the average width of the prediction band along with coverage guarantee, called Universally Trainable Optimal Predictive Intervals Aggregation (UTOPIA). The method also allows us to directly construct predictive bands based on elementary basis functions. Our approach is based on linear or convex programming which is easy to implement. All of our proposed methodologies are supported by theoretical guarantees on the coverage probability and optimal average length, which are detailed in this paper. The effectiveness of our approach is convincingly demonstrated by applying it to synthetic data and two real datasets on finance and macroeconomics.

9:30 **A VAE-based Framework for Learning Multi-Level Neural Granger-Causal Connectivity**
George Michailidis

Abstract: Granger causality has been widely used in various application domains to capture lead-lag relationships amongst the components of complex dynamical systems, and the focus in extant literature has been on a single dynamical system. In certain applications in macroeconomics and neuroscience, one has access to data from a collection of related such systems, wherein the modeling task of interest is to extract the shared common structure that is embedded across them, as well as to identify the idiosyncrasies within individual ones. This paper introduces a Variational Autoencoder (VAE) based framework that jointly learns Granger-causal relationships amongst components in a collection of related-yet-heterogeneous dynamical systems, and handles the aforementioned task in a principled way. The performance of the proposed framework is evaluated on several synthetic data settings and bench marked against existing approaches designed for individual system learning. The method is further illustrated on a real dataset involving time series data from a neurophysiological experiment and produces interpretable results.

10:00 **Regression discontinuity design with explained score**
Moulinat Banerjee, Debarghya Mukherjee, *Ya'acov Ritov*

Abstract: In non-randomized treatment effect models treatment is assigned to a unit based on some score, e.g., scholarship is allocated based on the score obtained in a merit-test or antihypertensive treatments are allocated based on blood pressure level. In this paper, we present a new model coined SCENTS: Score Explained Non-Randomized Treatment Systems, that utilizes the dependency of the score on the explanatory variables to permit an efficient estimation. We derived an estimator of the treatment effect which is sqrt-n consistent, asymptotically normal, and achieves emiparametric efficiency under normal errors. The analysis is extended to ultra-high dimensional vector of covariates, where a sqrt-n consistent and asymptotically normal debiased estimator is purposed. We analyze two real data sets via our method and compare our results with those obtained by using previous approaches like egression discontinuity design. Some possible extensions are discussed.

Advances in financial econometrics
Organizer: Genaro Sucarrat
Chair: Genaro Sucarrat
Room: Sala Polivalente 1.3

9:00 **Testing the zero-process of intraday financial returns for non-stationary periodicity**
Ovidijus Stauskas, *Genaro Sucarrat*

Abstract: Recent studies show that the zero-process of observed intraday financial returns is frequently characterised by non-stationary periodicity. As liquidity varies across the trading day, not only does unconditional volatility change, but also the unconditional zero-probability. While scaling returns by the time-varying intraday volatility may stabilise volatility, it does not make the zero-process of scaled returns stationary. This invalidates standard methods of risk estimation, and recent studies document that the use of such invalid methods can have major effects on risk estimates. Formal tests for non-stationary periodicity in the zero-process can therefore be of great value in guiding the choice of a suitable risk estimation procedure. Despite this, little attention has been devoted to the derivation of such tests. Here, we help filling this gap by developing user-friendly yet flexible and powerful tests that hold under mild assumptions. Next, an empirical study reveals that intraday financial returns are widely characterised by non-stationary periodicity in the zeroprocess. This has important and potentially wide-ranging implications for future research.

9:30 **Detection of breaks in weak location time series models with quasi-Fisher scores**
Christian Francq, Lorenzo Trapani, Jean-Michel Zakoian

Abstract: Based on Godambe's theory of estimating functions, we propose a class of cumulative sum, CUSUM, statistics to detect breaks in the dynamics of time series under weak assumptions. First, we assume a parametric form for the conditional mean, but make no specific assumption about the data-generating process (DGP) or even about the other conditional moments. The CUSUM statistics we consider depend on a sequence of weights that influence their asymptotic accuracy. Data-driven procedures are proposed for the optimal choice of the sequence of weights, in the sense of Godambe. We also propose modified versions of the tests that allow to detect breaks in the dynamics even when the conditional mean is misspecified. Our results are illustrated using Monte Carlo experiments and real financial data.

10:00 **Finite moments testing in a general class of nonlinear time series models**
Jean-Michel Zakoian, Christian Francq

Abstract: We investigate the problem of testing the finiteness of moments for a class of semi-parametric time series encompassing many commonly used specifications. The existence of positive-power moments of the strictly stationary solution is characterized by the Moment Determining Function (MDF) of the model, which depends on the parameter driving the dynamics and on the distribution of the innovations. We establish the asymptotic distribution of the empirical MDF, from which tests of moments are deduced. Alternative tests relying on the estimation of the Maximal Moment Exponent (MME) are studied. Power comparisons based on local alternatives and the Bahadur approach are proposed. We provide an illustration on real financial data and show that semi-parametric estimation of the MME offers an interesting alternative to Hill's nonparametric estimator of the tail index.

10:30 **Quantifying Uncertainty under Local Instability: a Dynamic Conformal approch to Electricity Price Forecasting**
Alessandro Giovannelli, Tommaso Proietti, Andrea Cerasa, Fany Nan

Abstract: This paper introduces a novel methodology, Dynamic Conformal Prediction (DPC), for the construction of prediction intervals with improved coverage in the presence of parameter instability. DCP allows the adaptation of both estimation and calibration windows. Through a Monte Carlo analysis, we demonstrate the effectiveness and validity of our procedure. In particular, using several data generating processes that differently account for parameter instability, we highlight the ability of our procedure to generate forecast intervals with a coverage that is close to the nominal one. Then, DCP will be employed for constructing prediction intervals for the time series of the single national price of the Italian Electricity Spot market in the short run, i.e., for forecast horizons that are not larger than 14 days ahead. The forecasting methods used encompass different assumptions (reduced forms, structural decompositions, nonlinearity) and degrees of mean reversion.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk 4

Chair: Jeffrey Racine
Room: Grande Auditorio

11:30 **Bootstrapping out-of-sample predictability tests with real-time data**
Silvia Gonçalves

Abstract: In this paper we develop a block bootstrap approach to out-of-sample inference when real-time data are used to produce forecasts. In particular, we establish its first-order asymptotic validity for West-type (1996) tests of predictive ability in the presence of regular data revisions. This allows the user to conduct asymptotically valid inference without having to estimate the asymptotic variances derived in Clark and McCracken's (2009) extension of West (1996) when data are subject to revision. Monte Carlo experiments indicate that the bootstrap can provide satisfactory finite sample size and power even in modest sample sizes. We conclude with an application to inflation forecasting that adapts the results in Ang et al. (2007) to the presence of real-time data. This is joint work with Michael McCracken and Yongxu Yao. [1] Ang, A., Bekaert, G., and M. Wei (2007), “Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?,” Journal of Monetary Economics, 54, 1163-1212. [2] Clark, T.E., and M.W. McCracken (2009), “Tests of Equal Predictive Ability with Real-Time Data,” Journal of Business and Economic Statistics 27, 441-54. [3] West, K.D. (1996), “Asymptotic Inference about Predictive Ability,” Econometrica, 64, 1067-1084.

12:30 - 13:30 Contributed 5

Biostatistics

Chair: M. Dolores Ruiz-Medina
Room: Sala Polivalente 1.1

12:30 **Sample Size Planning for the Wilcoxon-Mann-Whitney Test with Dependent Replicates**
Erin Sprünken, Frank Konietschke

Abstract: Every experiment should begin with careful sample size planning. If sample sizes are too small, procedures may neither control the type 1 error rate nor detect relevant effects, if they are too large, clinically irrelevant effects may be detected. Any trial should involve the minimum number of subjects required to detect a relevant effect with a given power. Typically, the effect of interest relies on the underlying distributions and data structure. Means and similar effects are not always appropriate to describe differences between the distributions, especially in case of skewed or ordinal data. In these situations, purely nonparametric Mann-Whitney effects can be used to describe treatment effects and are often used in statistical practice. However, the Wilcoxon-Mann-Whitney test does not allow for possibly dependent replicates, which often occur in various experiments in (pre-)clinical and socioeconomic trials. For instance, classes in schools, households with several members or patients providing two brain hemispheres in neurosciences are common examples of such clustered data. Ignoring this special data structure in the planning phase of a trial could lead to a severe over- or underestimation of the necessary sample size. Recently in a pioneering work, Rosner and Glynn (2011) proposed a sample size formula to detect a shift-alternative in the aforementioned model. However, formulating the effect of interest in terms of a shift-effect is only meaningful in very restricted models. Therefore, we aim to develop a sample size formula for the Wilcoxon-Mann-Whitney test with dependent replicates under mild assumptions in general scenarios. Extensive simulation studies demonstrate that our method estimates the sample size accurately in various scenarios and with various underlying distributions.

12:50 **Regression analysis for infectious disease modelling**
Chengyuan Lu, Jelle Goeman, Hein Putter, Mar Rodriguez Girondo

Abstract: Mathematical modeling of pandemic infectious diseases, such as the susceptible-infectious-recovered (SIR) model and its variations, often originates from structured models defined by differential equations. Despite the contagious rate varies with time and needs to be identified with time-to-event data, survival analysis techniques have been somewhat overlooked in this domain. This work aims to bridge this gap by establishing a connection between the SIR model and the additive hazards survival model. The additive hazards model is characterized by its purely non-parametric nature and is defined in terms of additive hazards and time-varying regression coefficients. We illustrate, theoretically, that these two model types are equivalent under ideal conditions. Specifically, the time-varying contagious rate inherent in the SIR model can be effectively derived through the time-varying coefficients from an additive hazards model. As a result, we propose a novel perspective on the SIR model, showing that its time-varying parameters can be identified through non-parametric regression models derived from survival analysis. To evaluate the performance of our innovative approach, we apply it to real-world data sourced from the Covid-19 pandemic in The Netherlands. Our research offers a new perspective on the comprehension and application of epidemiological models in the realm of public health.

13:10 **Including time-dependent variables in ROC curve analyses**
Arís Fanjul-Hevia, Juan Carlos Pardo-Fernandez, Wenceslao González-Manteiga

Abstract: Whenever there is a classification problem in which the aim is to differentiate two populations (such as healthy and diseased patients in a medical study), the concepts of sensitivity and specificity can be computed to assess the error measurements involved in the decision process. The Receiver Operating Characteristic (ROC) curve is a statistical tool that combines those notions to evaluate the discriminatory capability of the diagnostic marker under study. In cases where more than one method for diagnosing a certain disease is available, the comparison of the corresponding ROC curves is often used for comparing their accuracy of diagnosis. Apart from the diagnostic markers involved in the classification rule, in practice it is usual to have other covariates at our disposal. It is important take such information into consideration, as they may have an effect on the discriminatory capability of the diagnostic methods. There are several ways for incorporating the covariates to an ROC curve analysis, the main ones being the conditional ROC curves and the covariate-adjusted ROC curves. Several methodologies can be found in the literature for estimating those curves, some focused on a direct approach and others that integrate the covariate effect via induced regression models. In cases where the diagnostic marker or the covariates at hand are time related (for example, when dealing with longitudinal or functional data), these methodologies have to be adapted for the particularities of this type of data. The aim of this work is to explore the alternatives that exist in the literature to include the time dependent variables in the ROC curve analysis, particularly when the study is focused on the comparison of different diagnostic markers.

Nonparametric econometrics 2

Chair: Christian Francq
Room: Grande Auditorio

12:30 **Revisiting Localized Technical Change: A Nonparametric Instrumental Regression Approach with Mixed-Type Endogenous Regressors**
Davide Golinelli, Antonio Musolesi

Abstract: This paper explores localized technical change by investigating the impact of binary innovation variables on firms' production technology, challenging the notion of Hicks neutrality. It contributes to the literature on the econometrics of productivity and innovation by providing an empirical analysis of a nonparametric production function model. More specifically, we opt for a semiparametric partial linear specification. The parametric component is introduced in order to improve the modularity of the specification. In our case it allows to introduce sectors in estimating the production function without facing the curse of dimensionality problem. Since some of the inputs can be endogenous, we propose an alternative control function procedure designed to address mixed-type endogenous covariates combining local likelihood logit estimator and marginal integration technique. In order to assess the performance of the proposed estimator in finite samples, an extensive Monte Carlo simulation is performed. The Monte Carlo simulation showed that the estimator performs well under different level of endogeneity. Using data from Italian firms, the empirical findings from this study reveal two significant points: i) the proposed nonparametric estimator provides better predictive performance compared to alternative estimators, and ii) the proposed approach uncovers relevant patterns and localized effects that would remain undetected using traditional parametric methods.

12:50 **One-step smoothing splines instrumental regression**
Elia Lapenta, Jad Beyhum, Pascal Lavergne

Abstract: We extend nonparametric regression smoothing splines to a context where there is endogeneity and instrumental variables are available. Our estimator has several characteristics that should make it appealing for empirical work. First, our approach is particularly attractive because it is one-step. Two-step procedures typically lead to theoretical and practical issues: one may need to estimate in a first step an object that may be more complex than the final object of interest; and first-stage estimation typically affects the second-stage small sample and asymptotic properties. Second, a key benefit of the one- step nature of our estimator is that it depends upon one regularization parameter only. In existing two-step methods, each stage relies on a particular choice of a smoothing or regularization parameter, whose fine-tuning may be difficult in practice, while affecting the final results. In some methods, a third parameter is introduced to deal with the ill-posed nature of the inverse problem. To choose our unique regularization parameter, we devise a practical cross-validation method that yields good performances in simulations. Third, by contrast to previous approaches based on series or kernel methods, our estimator is a natural generalization of the popular smoothing splines estimator. The appeal of splines lies in their simplicity together with their excellent approximations of smooth functions. Fourth, due to its spline nature, our estimator is computationally simple, and a closed-form expression is easily obtained for the spline coefficients. Fifth, as an additional advantage, one obtains straightforward estimators of derivatives. We derive uniform rates of the convergence for the estimator and its first derivative. We also address the issue of imposing monotonicity in estimation. Simulations confirm the good performances of our estimator compared to two-step procedures. Our method yields economically sensible results when used to estimate Engel curves.

13:10 **The boosted Hodrick-Prescott filter, penalized least squares, and Bernstein polynomials**
Keith Knight

Abstract: The Hodrick-Prescott filter is commonly used in economics as a means of decomposing a time series into trend and cyclical components. Phillips and Shi (2019) proposed a modification of the Hodrick-Prescott filter, the boosted Hodrick-Prescott filter, which uses the idea of twicing (or boosting) to improve the behaviour of the Hodrick-Prescott filter. In this paper, we analyze some of the properties of the boosted Hodrick-Prescott filter by expressing it as the minimizer of a penalized least squares objective function that can also be approximated in the frequency domain. We also discuss an alternative modification of the Hodrick-Prescott filter using Bernstein polynomials. This approach, which we call the Hodrick-Prescott-Bernstein filter, can be tuned much more easily than the boosted Hodrick-Prescott filter.

Nonparametric inference and estimation 2
Chair: Eduardo García-Portugués
Room: Pequeno Auditorio

12:30 **Evaluating Randomness Assumption: A Novel Graph Theoretic Approach for Linear and Circular Data**
Shriya Gehlot, Arnab Kumar Laha

Abstract: Randomness is a fundamental assumption in various critical statistical theories in both linear and circular contexts. Therefore, assessing randomness in data holds pivotal significance in the practical application of these theories. However, the existing literature on randomness tests faces two significant limitations: (a) the availability of tests to assess randomness in linear data across diverse scenarios is limited, and (b) existing tests could not be applied to the circular data. This paper introduces a new approach to developing two non-parametric tests for linear and circular data. In circular data analysis, the existing literature predominantly focuses on testing the uniformity of the data, yet it lacks a comprehensive test designed to evaluate randomness within circular datasets. To address this gap, we introduce a new concept of Random Circular Arc Graphs (RCAG) in line with Random Interval Graphs (RIG). We explore various properties of the RCAGs, including edge probability, vertex degree distribution, maximum and minimum degrees, and the presence of Hamiltonian cycles. We also establish that, like RIG, these properties of the RCAG are independent of the choice of distribution of observations and depend solely on the independence of the data. We use this idea for RIG and RCAG to develop our tests to check randomness in the linear and circular data, respectively. Thus, our tests work for a wide range of application setups. We validate the efficacy of our tests through various simulation experiments. For linear data, we demonstrate that our test outperforms most of the standard parametric and nonparametric tests available in the literature, including the Runs test. Similarly, we substantiate the effectiveness of our tests for circular data through extensive simulations and real-world applications.

12:50 **Optimal non-parametric estimation of distribution functions, convergence rates for Fourier inversion theorems and applications**
Carlos Martins-Filho

Abstract: This paper contains two sets of results. First, we consider the broad class of kernel based non-parametric estimators of an unrestricted distribution function F proposed by Mynbaev, Martins-Filho and Henderson (2022). We develop improved lower and upper bounds for the bias of their estimators at points of continuity of F as well as for jumps of F. In addition, we obtain necessary and sufficient conditions for asymptotic unbiasedness of estimators in the class. Second, we provide new Fourier inversion theorems with rates of convergence and use them to obtain convergence results for new classes of deconvolution estimators for distribution functions and their jumps under measurement error.

13:10 **Partial identification for a wide class of event time models in the presence of dependent censoring**
Ilias Willems, Ingrid Van Keilegom, Jad Beyhum

Abstract: Understanding event times is crucial in various fields, yet it often involves censoring mechanisms which challenge traditional approaches. In the current literature, many models require stringent assumptions in order to guarantee valid inference in these settings, especially when the censoring is dependent on the event of interest. In this research, we refrain from such an approach by proposing a general model for the latent event times which makes minimal assumptions about the censoring mechanism. More specifically, we model the distribution of the possibly unobserved event time conditional on the covariates using a parametric model with a known link function. This link function can take many forms and can be specified by the practitioner. Because the event time is not always observed, it is in general not possible to uniquely identify our model. Therefore, we propose to partially identify it. A partially identified model is one in which the parameters of interest are not uniquely determined by the distribution of the observed data and the maintained assumptions. In this paper, we achieve partial identification by exploiting Peterson's bounds on the conditional distribution of the event time and casting the problem into one defined by unconditional moment restrictions, which allows us to rely on a vast basis of established theory. Because of the flexibility in choosing the link function, our formulation includes many commonly used models in survival analysis, like the parametric Cox model or the accelerated failure time model. Moreover, we show how our formulation permits us to study cure fractions in an assumption-lean manner. A simulation study illustrates the finite sample performance of the proposed approach.

13:30 - 14:30 Lunch & Farewell