EMS 2019

Program and Book of Abstracts

32nd Edition of the
European **M**eeting of **S**tatisticians

Palermo, July 22-26

Bernoulli Society
for Mathematical Statistics
and Probability

Title: EMS 2019 - Program and Book of Abstracts
Program and Book of Abstracts of the 32nd edition of the European Meeting of Statisticians
Palermo (Italy), 22–26 July 2019
http://www.ems2019.palermo.it

Edited by: Angelo M. Mineo and Luigi Augugliaro

Year: 2019

Programme Committee

Lutz Dümbgen	(Chair)
Anne Gégout-Petit	(ERC Programme coordinator)
Angelo Mineo	(Local organizing committee representative)
Holger Dette	
Ingrid Glad	
Irène Gijbels	
Omiros Papaspiliopoulos	
Qiwei Yao	

Local Organizing Committee

Chair Angelo Mineo	(Chair)	angelo.mineo@unipa.it
Co-Chair Luigi Augugliaro	(Co-Chair)	luigi.augugliaro@unipa.it
Antonino Abbruzzo		antonino.abbruzzo@unipa.it
Marcello Chiodi		marcello.chiodi@unipa.it
Francesca Di Salvo		francesca.disalvo@unipa.it
Francesco Mola		mola@unica.it
Antonella Plaia		antonella.plaia@unipa.it
Antonella Ruggieri		antonella.ruggieri@unipa.it
Nicola Torelli		nicola.torelli@deams.units.it

Contents

Forewords from the Chair of the EMS 2019

The European Meeting of Statisticians (EMS) is the main conference in statistics and probability in Europe. The very first EMS meeting was held in Dublin in 1962, and the 32nd EMS - EMS 2019 - takes place in Palermo, the capital of Sicily (Italy). The meeting is co-sponsored by the European Regional Committee of the Bernoulli Society, the University of Palermo, the University of Cagliari, the Polytechnic University of Turin, the University of Trieste and the Italian Statistical Society (SIS).

The meeting is a collective effort and many individuals and organizations have contributed. The Programme Committee, the Local Organizing Committee, the organizers of the Invited Sessions, the local hosting university and volunteers, all of them have played to a substantial role in the organization of the meeting. We warmly acknowledge their work and the support of all the institutions and organizations that have sponsored the meeting.

The meeting is hosted in Palermo. Palermo is a city rich in monuments that testify its centuries-old history. Some monuments are of particular interest: the Norman Palace (Palazzo dei Normanni), built around 1100, which is the oldest Royal residence in Europe; inside the Norman Palace there is also the famous Cappella Palatina; the Cathedral, built in its present form in 1185 on the basis of an older church dated from the 4th century; the Martorana Church (Chiesa della Martorana) of Arabic style built in 1143; the Teatro Massimo, built in 1875, which is the largest opera house in Italy and one of the largest in Europe. Palermo is also famous for its traditional food. There are also many good restaurants for vegan and vegetarian. Then, all the partecipants can enjoy the city.

The 32nd edition of the European Meeting of Statisticians put together recent advances and trends in several fields of statistics and probability theory. With 6 plenary speakers, 25 invited sessions, 20 topic-contributed sessions, 35 contributed sessions and a poster session, and with about 450 participants, EMS19 is the perfect place to exchange ideas and promote collaboration between researchers from all the World.

We wish you a productive, stimulating conference and a memorable stay in Palermo.

Angelo M. Mineo

Outline

		Monday July 22			Tuesday July 23			Wednesday July 24		
08:00	08:30	**Registration**								
08:30	09:00	**Welcome Words** **Aula Magna - Building 13**								
09:00	09:30	**Opening Lecture**			**Special Invited Lecture**			**Forum Lecture I**		
		Speaker	J. Rousseau		Speaker	G. Allen		Speaker	V. Panaretos	
09:30	10:00	Chair	A. M. Mineo		Chair	Q. Yao		Chair	I. Gijbels	
		Room	Aula Magna - Building 13		Room	Aula Magna - Building 13		Room	Aula Magna - Building 13	
10:00	10:30	**Coffee Breack**			**Coffee Breack**			**Coffee Breack**		
		Code	**Room**	**Organizer**	**Code**	**Room**	**Organizer**	**Code**	**Room**	**Organizer**
10:30	11:00	IS-02	11	L. Aue	IS-18	12	W. Polonik	IS-05	12	I. Castillo
		IS-06	9	N. Chopin	IS-24	11	C. Varin	IS-13	11	C. Ley
11:00	11:30	TCS-01	7	S. Meintanis	IS-25	9	M. van den Wiel	IS-14	9	R. Liu
		TCS-02	6	L. Augugliaro	TCS-09	7	A. Kume	TCS-14	7	K. Kamatani
11:30	12:00	TCS-03	5	Q.-M. Shao	TCS-10	5	S.-F. Huang	TCS-15	5	H. Masuda
		CS-01	12		CS-12	10		CS-20	10	
		CS-02	10		CS-13	8		CS-21	8	
12:00	12:30	CS-03	8		CS-14	6		CS-22	6	
		CS-04	Seminari B		CS-15	Seminari B		CS-23	Seminari B	
12:30	14:00	**Lunch Break**			**Lunch Break**			**Lunch Break**		
		Code	**Room**	**Organizer**	**Code**	**Room**	**Organizer**			
14:00	14:30	IS-04	12	F. Bunea	IS-03	11	F. Bach			
		IS-11	11	R. Koenker	IS-10	9	M. Hoffmann			
14:30	15:00	IS-22	9	G. Stupfler	TCS-11	7	D. Politis			
		TCS-04	7	R. V. Craiu	TCS-12	6	C. Klüppelberg			
15:00	15:30	CS-05	10		TCS-13	5	C. Améndola			
		CS-06	8		CS-16	12				
		CS-07	6		CS-17	10				
15:30	16:00	CS-08	Seminari B		CS-18	8				
					CS-19	Seminari B				
16:00	16:30	**Coffee Breack**			**Coffee Breack**					
		Code	**Room**	**Organizer**				**Excursion**		
16:30	17:00	IS-17	12	D. Paindaveine						
		IS-19	11	A. Särkkä	**Poster Session**					
17:00	17:30	IS-20	9	R. Samworth						
		TCS-06	7	R. Stelzer						
		TCS-07	6	A. Tartakovsky						
17:30	18:00	TCS-08	5	K. Bharath	**BS-EMS Lecture**					
		CS-09	10							
		CS-10	8		Speaker	A. van der Vaart				
18:00	18:30	CS-11	Seminari B		Chair	M. Gyllenberg				
					Room	Aula Magna - Building 13				
18:30	20:00									
20:00		**Welcome Reception**								

Time		Thursday July 25	Friday July 26
08:00	08:30		
08:30	09:00		
09:00	09:30	**Forum Lecture II** Speaker: V. Panaretos Chair: L. Dümbgen Room: Aula Magna - Building 13	Parallel sessions (see below)
09:30	10:00		
10:00	10:30	**Coffee Breack**	
10:30	11:00	Parallel sessions (see below)	
11:00	11:30		**Coffee Breack**
11:30	12:00		**Closing Lecture** Speaker: J. Lafferty Chair: A. Gegout-Petit Room: Aula Magna - Building 13
12:00	12:30		
12:30	14:00	**Lunch Break**	
14:00	14:30	Parallel sessions (see below)	
14:30	15:00		
15:00	15:30		
15:30	16:00		
16:00	16:30	**Coffee Breack**	
16:30	17:00	**Special Invited Lecture** Speaker: G. Blanchard Chair: I. Glad Room: Aula Magna - Building 13	
17:00	17:30		
17:30	18:00		
18:00	18:30		
18:30	20:00		
20:00		**Conference Dinner**	

Thursday July 25, 10:30 sessions

Code	Room	Organizer
IS-09	12	Y. Goude
IS-12	11	E. Kolaczyk
IS-16	9	J. Neslehova
TCS-16	7	A. R. Brazzale
TCS-17	5	K. Kamatani
CS-24	10	
CS-25	8	
CS-26	6	
CS-27	Seminari B	

Thursday July 25, 14:00 sessions

Code	Room	Organizer
IS-01	12	L. Aslett
IS-21	11	L. Sangalli
IS-23	9	Y. Swan
TCS-18	7	T.-H. Li
TCS-19	5	L. Martino
CS-28	10	
CS-29	8	
CS-30	6	
CS-31	Seminari B	

Friday July 26, 09:00 sessions

Code	Room	Organizer
IS-07	12	I. Eckley
IS-08	11	Y. Fan
IS-15	9	A. Munk
TCS-20	10	D. Edelmann
TCS-05	Seminari B	M. Hadjikyriakou
CS-32	5	
CS-33	6	
CS-34	7	
CS-35	8	

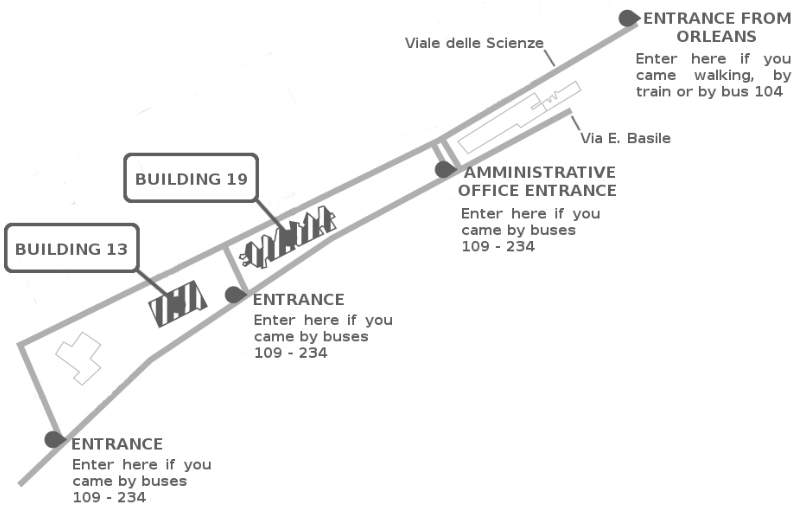
Legend

IS	Invited Session
TCS	Topic Contributed Session
CS	Contributed Session

General Information

Conference venue

The Conference is held at the Conference and Didactic Centre of the University of Palermo, inside the University Campus, located in Viale delle Scienze (a parallel of Via Ernesto Basile), in buildings 19 (edificio 19) and 13 (edificio 13) for the general talks and for plenary sessions, respectively. Despite of the numbering, the two buildings are one in front of the other. Below there is a map of the Campus, with the location of buildings 19 and 13.



Invited (IS), Topic-Contributed (TCS) and Contributed (CS) Sessions are held in Sections B and C of Building 19, rooms 5 - 12 at first floor (see Figure 1). Poster session is held at the second floor in Section B.

Registration

The EMS19 staff will be at the registration desk placed in section B of Building 19. Registration will start at 8:00 a.m. on Monday 22 and will be open for the whole time of the conference.

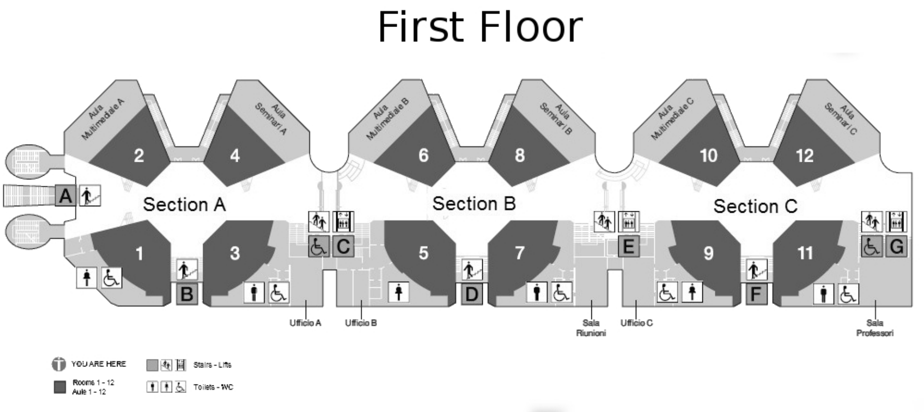


Figure 1: Map of the first floor, building 19

Presentation instructions

The rooms are equipped with a PC and a computer projector. Presenters must provide the files with the presentation to the session chair at the beginning of the session. Files must be in PDF (Adobe) or PPT (PowerPoint) format on a USB memory stick. Speakers in Invited Sessions have 40 minutes for presentation, including questions and comments. Speakers in Topic-Contributed and Contributed Sessions have 20 minutes for presentation, including questions and comments. Speakers are invited to bring their presentations and upload them to the PC in the session room no later than 15 minutes before the begining of your session.

Chairs are kindly requested to keep strictly to the schedule. Paper should be presented in the order they are listed in the program for the convenience of attendees, who may wish to go to other rooms to hear particular speakers talk. In case of absence of the presenter, please use extra-time for a break or for a discussion, in order to remain on schedule.

Technicians from the conference organization will be available during the conference. Please, refer to the registration desk in case of problems.

Poster instructions

Each author will be provided with one poster board 95 cm across (3 ft 1.4 inches) and 114 cm high (3 ft 8.9 inches). We recommend the standard poster A0 size (portrait). Posters are to be set-up during the conference at any time before the beginning of the poster session on Tuesday 23 at 16:30-17:30 and presenters must be available during the poster session. Posters must be taken off from the poster board before the end of the conference.

Social dinner

The Conference Organizing Secretariat have organized a social dinner that will take place on Thursday 25.

The social dinner event will take place at “Castello a Mare” (Castle on the sea), one of the most suggestive areas of the city and will be a great chance to network with the Conference participants and speakers and taste some great dishes of the Sicilian cuisine.

Internet connection

The entire University Campus is covered by the Eduroam network. In “Aula Multimediale C” there are computers connected to Internet for participants’ use.

Program

Monday, July 22

08:00-08:45 Registration

08:45-09:00 Welcome Words

Aula Magna - Building 13

09:00-10:00 Plenary Session

PL-01 Opening lecture

Plenary, Aula Magna - Building 13
Chair: Angelo M. Mineo

Judith Rousseau	Bayesian measures of uncertainty

10:30-12:30 Invited and Contributed Sessions

IS-02 Time series analysis for complex data

Invited session, Room 11 (Aula 11)
Organizer: Alexander Aue
Chair: Alexander Aue

Giuseppe Cavaliere	Bootstrapping non-stationary stochastic volatility
Rainer Dahlhaus	Towards a general theory for non-linear locally stationary processes
Richard A. Davis	Noncausal vector AR processes with application to economic time series
Suhasini Subba Rao	Feature estimation and testing for linear regression with time series regressors

IS-06 Bayesian computation

Invited session, Room 9 (Aula 9)
Organizer: Nicolas Chopin
Chair: Nicolas Chopin

Simon Barthelmé	Kernel matrices in the flat limit
Nicolas Chopin	Adaptive tuning of Hamiltonian Monte Carlo within sequential Monte Carlo
Giacomo Zanella	On the robustness of gradient-based MCMC

TCS-01 Goodness-of-fit and change-point methods

Topic-contributed session, Room 7 (Aula 7)
Organizer: Simos Meintanis
Chair: Simos Meintanis

Dimitrios Bagkavos	A smooth goodness-of-fit test for densities
Bojana Milošević	New consistent characterization based goodness-of-fit tests
Marko Obradović	Comparison of symmetry tests in i.i.d. and non i.i.d. setting
Charl Pretorius	Detection of changes in panel data models with stationary regressors
Simos Meintanis	Change-Point detection with multivariate data: two-sample situations

TCS-02 Advanced statistical methods for high-dimensional and complex data

Topic-contributed session, Room 6 (Aula 6)
Organizer: Luigi Augugliaro
Chair: Luigi Augugliaro

Antonino Abbruzzo	Probabilistic networks for GPS data analysis
Giada Adelfio	Weighted local second-order statistics for complex spatio-temporal point processes
Francesca Martella	Biclustering for multivariate longitudinal data
Gianluca Sottile	An extension of the censored gaussian lasso estimator

TCS-03 Advances in asymptotic theory

Topic-contributed session, Room 5 (Aula 5)

Organizer: Qi-Man Shao
Chair: Qi-Man Shao

Andreas Anastasiou	Bounds for the asymptotic distribution of the likelihood ratio
Bhaswar Bhattacharya	Detection thresholds for non-parametric tests based on geometric graphs
Xiao Fang	Wasserstein-2 bounds in normal approximation under local dependence
Matthias Schulte	Normal approximation of stabilising functionals
Qi-Man Shao	Recent progress of self-normalized limit theory

CS-01 Recent advances in density estimation

Contributed session, Room 12 (Aula 12)
Chair: Tullia Padellini

Zahra Amini Farsani	Optimization algorithm for parameter estimation in cell biology
Magnus Ekström	A class of asymptotically efficient estimators based on sample spacings
Federico Ferraccioli	Nonparametric likelihood density estimation on bounded domains
Ruey-Ching Hwang	Predicting LGD distributions with mixed continuous and discrete ordinal outcomes
Martin Kroll	Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids
Tullia Padellini	Persistent homology for kernel density exploration

CS-02 Functional data analysis

Contributed session, Room 10 (Aula 10)
Chair: Andrej Srakar

Melanie Birke	Simultaneous confidence bands for the covariance kernel of Banach space valued functional data
Alessia Caponera	Asymptotics for spherical autoregressions
Alessia Pini	Measuring reliability of functional data: an application to human movements
Tomas Rubin	Sparsely observed functional time series: estimation and prediction
Maria Skupień	Big data and functional approach to signal analysis
Andrej Srakar	Program evaluation and causal inference for distributional and functional data: estimation of the effects of retirement on health outcomes

CS-03 New approaches for analyzing high-dimensional data

Contributed session, Room 8 (Aula 8)
Chair: Jing Zhou

Vera Djordjilović	Global test for high-dimensional mediation: testing groups of potential mediators
Zhaoyuan Li	A precise framework for testing weak instruments in high-dimensional instrumental variables regression
Myrto Limnios	Empirical maximization of R-statistics in the two-sample problem and nonparametric homogeneity tests in high dimension
Yiming Liu	Covariance estimation of matrix-variates through iterative process
Jing Zhou	Weight choice for composite and model-averaged estimation

CS-04 Bayesian inference

Contributed session, Seminari B
Chair: Cristina Gutiérrez Pérez

Themistoklis Botsas	Bayesian approach to deconvolution in Well test analysis
Annalisa Cerquetti	Shannon entropy estimation via discovery rates: linking frequentist and Bayesian solutions
Alberto J. Coca	Nonparametric posterior contraction rates for discretely observed compound Poisson processes
Fabio Corradi	Relative privacy threats and learning from anonymized data
Matteo Giordano	Bernstein-von Mises teorems and uncertainty quantification for linear inverse problems
Cristina Gutiérrez Pérez	Bayesian inference for a Y-linked pedigree with non-syndromic hearing impairment

14:00-16:00 Invited and Contributed Sessions

IS-04 High dimensional inference in structured models

Invited session, Room 12 (Aula 12)
Organizer: Florentina Bunea
Chair: Florentina Bunea

Boaz Nadler	Robust sparse covariance estimation by thresholding Tyler's M-estimator
Philippe Rigollet	Minimax rates of estimation for smooth optimal transport maps
Marten Wegkamp	Essential regression

IS-11 Empirical Bayes methods

Invited session, Room 11 (Aula 11)
Organizer: Roger Koenker
Chair: Roger Koenker

Jiaying Gu	Nonparametric empirical Bayes methods for ranking with longitudinal data
Yihong Wu	Optimal estimation of Gaussian mixtures via denoised method of moments
Cun-Hui Zhang	Second order Stein: SURE for SURE and other applications

IS-22 Statistical aspects of extreme value analysis

Invited session, Room 9 (Aula 9)
Organizer: Gilles Stupfler
Chair: Gilles Stupfler

Marco Oesting	Estimation of the spectral measure of regularly varying random vectors
Simone Padoan	Nonparametric Bayesian estimation of the extremal dependence
Holger Rootzén	Is climate change making extreme rains more frequent, or bigger, or more dangerous?

TCS-04 Statistics in biosciences

Topic-contributed session, Room 7 (Aula 7)
Organizer: Radu V. Craiu
Chair: Radu V. Craiu

Radu Craiu	Optimal designs in post-genome-wide association two-phase sampling studies
Giancarlo Manzi	Biases in bias elicitation
Lei Sun	The X factor: a robust and powerful approach to X-chromosome-inclusive whole-genome association studies
Ronghui Xu	Applying additive hazards model to learn from electronic health data

CS-05 New developments in Bayesian analysis

Contributed session, Room 10 (Aula 10)
Chair: Hector Zarate

Ingvild Helgoy	A noise-robust fast sparse Bayesian lerning model
Kostas Kalogeropoulos	On Bayesian structural equation modelling
Tatyana Krivobokova	Adaptive non-parametric estimation of mean and autocovariance in regression with dependent errors
Suleyman Ozekici	Bayesian analysis of hidden markov models
Panayiota Touloupou	Scalable inference for epidemic models with individual level data
Hector Zarate	Semiparametric modeling of the mean, variance and scale parameters in skew normal regression models: a Bayesian perspective

CS-06 New developments in nonparametric regression

Contributed session, Room 8 (Aula 8)
Chair: Alexander Kreiß

Cécile Adam	Flexible modelling and expectile regression
Kin Yap Cheung	High-dimensional local linear regression under sparsity and convex losses
Maarten Jansen	Multiscale local polynomial estimation from highly irregular data
Prashant Jha	A new graphical device and related tests for the shape of nonparametric regression function
Nick Kloodt	Two goodness of fit tests in nonparametric transformation models

Alexander Kreiß	Statistical inference on networks: Correlation bounds, mixing and m-dependence under random, time-varying network distances with applications to a dynamic network model based on counting processes

CS-07 Recent advances in nonparametric regression

Contributed session, Room 6 (Aula 6)
Chair: Andrej Srakar

Mercedes Conde Amboage	Testing quantile regression models when the response variable is right censored
Rezaul Karim	Semiparametric quantile regression using quantile-based asymmetric family of densities
Alexandre Moesching	Shape and order constraints in nonparametric regression
Matti Pajari	Probabilistic comparison of quantile estimators for continuous random variables
Juan C. Pardo-Fernandez	Testing for superiority between two variance functions
Andrej Srakar	Wavelet regression estimator for compositional data

CS-08 New trends in survival analysis and applied statistics

Contributed session, Seminari B
Chair: Grzegorz Wyłupek

Tom Britton	Statistical challenges when analysing emerging epidemic outbreaks
Negera W. Deresa	Bivariate parametric model for survival data under dependent censoring
Wenqing He	Perturbed-variance based null hypothesis tests with an application to Clayton models
Ayman Hijazy	Gamma process based models for disease progression
Chun Y. Lee	Testing for change-point in the covariate effects based on the Cox proportional hazards model
Grzegorz Wyłupek	Data-driven Kaplan-Meier one-sided two-sample tests

16:30-18:30 Invited and Contributed Sessions

IS-17 Optimality in high-dimensional statistics

Invited session, Room 12 (Aula 12)
Organizer: Davy Paindaveine
Chair: Davy Paindaveine

Olivier Collier	Polynomial-time estimation of the mean
Alexei Onatski	Local asymptotic normality in high-dimensional spiked models
Thomas Verdebout	Detecting the direction of high-dimensional spherical signals

IS-19 Spatial and space time modelling

Invited session, Room 11 (Aula 11)
Organizer: Aila Särkkä
Chair: Aila Särkkä

Anastassia Baxevani	Effective probability distributions for spatially dependent processes
Ute Hahn	Clusterwise marked spatial cluster point processes, with application to single molecule microscopy
Martin Schlather	Advanced models and methods for bivariate random fields

IS-20 Measuring and testing independence

Invited session, Room 9 (Aula 9)
Organizer: Richard J. Samworth
Chair: Richard J. Samworth

Jonas Peters	The impossibility of conditional independence testing and a way out
Niklas Pfister	Kernel-based tests for joint independence
Richard Samworth	Nonparametric independence testing via mutual information

TCS-06 Modelling and estimation of spatial and space time data

Topic-contributed session, Room 7 (Aula 7)
Organizer: Robert Stelzer
Chair: Robert Stelzer

David Berger	CARMA generalized processes and stochastic partial differential equations
Jürgen Kampf	Nonparametric estimation of the kernel function of stable moving average processes
Claudia Klueppelberg	Estimation of causal CARMA random fields
Bennet Ströh	Weak dependence of causal random fields and statistical applications
Qinwen Wang	Identification of the number of factors for factor modeling in high dimensional time series

TCS-07 Sequential methodologies and their applications

Topic-contributed session, Room 6 (Aula 6)
Organizer: Alexander Tartakovsky
Chair: Alexander Tartakovsky

Michael Baron	Change-point detection in a Markov random field
Olympia Hadjiliadis	Trends and trades
Serguei Pergamenchtchikov	Asymptotically optimal pointwise and minimax changepoint detection for general stochastic models with a composite post-change hypothesis
Yao Xie	Scan B-statistic for kernel change-point detection
Alexander Tartakovsky	Optimal change detection rules maximizing probability of detection and their application for efficient detection of near-earth space object tracks

TCS-08 Geometry and invariance in analysis of shape and functional data

Topic-contributed session, Room 5 (Aula 5)
Organizer: Karthik Bharath
Chair: Karthik Bharath

Huiling Le	Limiting behaviour of Frechet Means on manifolds
Anuj Srivastava	Shapes analysis of functional data
Alfred Kume	Some results on the linear combination of chi-squares with applications to inference in shape and directional statistics
Karthik Bharath	Invariant aspects of functional data with phase variation

CS-09 Spatial analysis: modeling

Contributed session, Room 10 (Aula 10)
Chair: Hanne Rognebakke

Nan-Jung Hsu	Forecast-emphasized principal component for spatial temporal data
Hsin-Cheng Huang	False discovery rates to detect signals from incomplete spatially aggregated data
Yen-Ning Huang	Multivariate spatial models for severe storm activities
Hans Karlsen	A Spatial ARMA-GARCH model
Aleksandar Kolev	Spatially explicit capture recapture as a self exciting point process
Hanne Rognebakke	Statistical space-time projections of wave heights in the North Atlantic

CS-10 Time-series analysis: modeling

Contributed session, Room 8 (Aula 8)
Chair: Martina Vittorietti

Annika Betken	Testing for a change in the tail index of long-memory stochastic volatility time series
Paolo E. Cardone	Four factor model in Italian equities market
Feiqing Huang	Gaussian quasi-maximum likelihood estimation for linear GARCH models
Božidar Popović	A bivariate binomial count time series with application to the number of rainfall days
Lena Reichmann	Autoregressive-type time series models with bounded support
Martina Vittorietti	2D real microstructures images: a great source of data

CS-11 Change point analysis

Contributed session, Seminari B
Chair: Michal Pesta

Alexander Dürre	Robust change-point detection in panel data
Farid El Ktaibi	Bootstrapping the empirical distribution of a stationary process with change-point
Matus Maciak	Structural breaks in nonparametric models via atomic pursuit methods
Pekka Malo	Non-parametric structural change detection in multivariate systems

Sévérien Nkurunziza	On some inferences in generalized Ornstein-Uhlenbeck processes with multiple change-points
Michal Pesta	Changepoint in non-stationary series without nuisance parameters

Tuesday, July 23

09:00-10:00 Plenary Session

PL-02 Special Invited Lecture

Plenary, Aula Magna - Building 13
Chair: Qiwey Yao

Genevera Allen	Data integration: data-driven discovery from diverse data sources

10:30-12:30 Invited and Contributed Sessions

IS-18 Statistical topological data analysis

Invited session, Room 12 (Aula 12)
Organizer: Wolfgang Polonik
Chair: Wolfgang Polonik

Eddie Aamari	Estimating the reach of a manifold
Claire Brécheteau	Robust shape inference from a sparse approximation of the Gaussian trimmed loglikelihood.
Johannes Krebs	Advances on the asymptotic normality of Betti numbers

IS-24 Likelihood extensions and modifications

Invited session, Room 11 (Aula 11)
Organizer: Cristiano Varin
Chair: Paolo Vidoni

David Firth	Quasi-likelihood for compositional data
Emil A. Stoltenberg	Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes
Grace Yi	Modeling and analysis of correlated data using pairwise likelihood

IS-25 Bayes and empirical Bayes for genomics

Invited session, Room 9 (Aula 9)

Organizer: Mark van de Wiel
Chair: Mark van de Wiel

Gwenaël G.R. Leday	Fast Bayesian inference in large graphical models
Francesco Claudio Stingo	Varying-sparsity regression models with application to cancer proteogenomics
Mirrelijn Van Nee	An Empirical Bayes approach to co-data learning in ridge models

TCS-09 Recent developments in statistics of manifold data

Topic-contributed session, Room 7 (Aula 7)
Organizer: Alfred Kume
Chair: Alfred Kume

Tomoaki Imoto	A method for constructing a circular distribution from a single valued analytic function
Kei Kobayashi	Statistical inference and data analysis on length metric spaces
Tomonari Sei	A weighted fshape model in computational anatomy
Katie Severn	Manifold valued data analysis of samples fo networks, with applications in corpus linguistics

TCS-10 Recent developments in statistical learning and time series analysis for complex data

Topic-contributed session, Room 5 (Aula 5)
Organizer: Shih-Feng Huang
Chair: Shih-Feng Huang

Chihhao Chang	Estimation of breakpoints for extended interval regression models
Meihui Guo	Financial time series analysis with unsupervised learning
Shih-Feng Huang	Classification of temporal data using dynamic time warping and compressed learning
Liang-Ching Lin	Modeling financial interval time series
Mong-Na Lo Huang	Recurrent neural network for short-term load forecasting with spline bases
Li-Hsien Sun	Estimation under copula-based Markov mixture normal models for time series data

CS-12 Modeling high dimensional data

Contributed session, Room 10 (Aula 10)

Chair: Jakob Söhl

Habib Ganjgahi	Scalable multivariate Bayesian model for multi-level data
Niels R. Hansen	Post-selection inference: risk estimation after data-driven model selection
Ker-Chau Li	A new multivariate analysis framework to investigate complex interaction patterns
Bastien Marquis	Image denoising using corrected information criterion and grouping
Jakob Söhl	Noise fit, estimation error and a sharpe information criterion

CS-13 Computer intensive methods

Contributed session, Room 8 (Aula 8)
Chair: Bambang Suprihatin

Alexander Braumann	Bootstrapping for impulse response function coefficients in generalized linear dynamic factor models
Rasmus Erlemann	Simulating from conditional distributions
Marta Galvani	Generalized Bayesian ensemble modeling: methodological and computational aspects
Valentina Mameli	Bootstrap group penalties for predicting molecular properties
Marepalli Rao	Data analysis in the environment of destructive samples: the case of spina bifida
Bambang Suprihatin	Asymptotic distribution of the bootstrap parameter of an $AR(p)$ model

CS-14 Recent advances in computationally intensive statistics

Contributed session, Room 6 (Aula 6)
Chair: Michael Weylandt

Apostolos Batsidis	Berkson's paradox and weighted distributions: an application to Alzheimer's disease
Ruggero Bellio	Point estimation based on confidence intervals
Kun-Lin Kuo	Efficient computation of pseudo-Gibbs distributions
Tabea Rebafka	Properties of the stochastic approximation EM algorithm with mini-batch sampling
Cecilia Viscardi	Weighted approximate Bayesian computation via Sanov's theorem
Michael Weylandt	Dynamic visualization and fast computation for convex clustering via algorithmic regularization

CS-15 Graphical models

Contributed session, Seminari B (Seminari B)
Chair: Eugen Pircalabelu

Melih Agraz	Gaussian graphical model under different model selection criteria for Lasso regression
Ezgi A. Demirci	A new steady-state modeling approach for protein-protein interaction networks
Andrea Lazzerini	Bayesian stochastic search for Ising chain graph models
Margot Matterne	The average conditional and partial Kendall's tau
Federica Onori	Experience-based food insecurity scales, a non-aggregative approach to synthesis of indicators
Eugen Pircalabelu	Community detection for probabilistic graphical models

14:00-16:00 Invited and Contributed Sessions

IS-03 Machine learning and optimization

Invited session, Room 11 (Aula 11)
Organizer: Francis Bach
Chair: Francis Bach

Quentin Berthet	Optimal transport methods in statistics and machine learning
Claire Boyer	On the structure of solutions of convex regularization: gradient TV minimization and co
Alessandro Rudi	Scaling up optimal kernel methods for large scale machine learning

IS-10 Statistical inference and PDE's

Invited session, Room 9 (Aula 9)
Organizer: Marc Hoffmann
Chair: Marc Hoffmann

Randolf Altmeyer	Nonparametric drift estimation for linear SPDEs from local measurements
Richard Nickl	Consistent Bayesian inference for some elliptic PDE models
Vincent Rivoirard	Nonparametric estimation for size-structured population of cells

TCS-11 Bootstrap and time series

Topic-contributed session, Room 7 (Aula 7)
Organizer: Dimitris Politis
Chair: Dimitris Politis

Carsten Jentsch	Bootstrapping Hill estimator and tail arrays sums for regularly varying time series
Jonas Krampe	Bootstrap based inference for sparse high-dimensional time series models
Michele La Rocca	NAR-Sieve bootstrap based on extreme learning machines
Marco Meyer	Extending the validity of frequency domain bootstrap methods to general stationary processes
Efstathios Paparoditis	Recent advances in bootstrapping functional time series
Dimitris Politis	Predictive inference for locally stationary time series

TCS-12 High frequency, continuous time and non-stationarity

Topic-contributed session, Room 6 (Aula 6)
Organizer: Claudia Klüppelberg
Chair: Claudia Klüppelberg

Dirk-Philip Brandes	Lévy driven continuous time moving average and strong mixing processes sampled at a renewal sequence
Thiago Do Rego Sousa	Method of moment based estimation for the multivariate COGARCH(1,1) processes
Mark Podolskij	On optimal estimation of random quantities associated with Levy processes
Daniel Rademacher	Asymptotic normality of integrated periodogram operators
Robert Stelzer	Continuous-time locally stationary time series models

TCS-13 Algebric statistics

Topic-contributed session, Room 5 (Aula 5)
Organizer: Carlos Améndola
Chair: Carlos Améndola

A. Grosdos Koutsoumpelias	Moments of local Dirac mixtures - algebra and applications in statistics
Orlando Marigliano	Discrete statistical models with rational maximum likelihood estimator
Fabio Rapallo	Circuits in experimental design
Elina Robeva	Maximum likelihood estimation for totally positive densities

Carlos Améndola	Autocovariance varieties of moving average random fields

CS-16 Advanced topic in nonparametric modeling

Contributed session, Room 12 (Aula 12)
Chair: Qiwei Yao

Vahe Avagyan	Stable IPW estimation for longitudinal studies
Beatrice Laurent	Aggregated tests of independence based on HSIC measures
Anouar Meynaoui	Aggregated tests of independence based on HSIC measures
Paula Saavedra-Nieves	Data-driven support estimation
Maria-Pia Victoria-Feser	Bias reduced simulation-based estimators in high dimensions
Qiwei Yao	Estimation of subgraph densities in noisy networks

CS-17 Robust and nonparametric methods

Contributed session, Room 10 (Aula 10)
Chair: Zbigniew S. Szewczak

Giacomo Francisci	A modification of the simplicial depth function
Stanislav Nagy	The halfspace depth characterization problem
Beatriz Pateiro-López	Rate-optimal estimators for the volume of a set
Serhan Sadikoglu	Misclassification-robust semiparametric estimation of single-index binary choice models
Adrien Saumard	Robust to outliers of median-of means
Zbigniew S. Szewczak	On martingale CLT for strictly stationary sequences

CS-18 Advanced topic in econometrics

Contributed session, Room 8 (Aula 8)
Chair: Riccardo Parviero

Tobias Fissler	Consistent scoring functions and murphy diagrams for set-valued measures of systemic risk
Vladimír Holý	Modeling discrete trade durations with excessive zeros
Paul Kvam	Using randomization testing to estimate the effects of gerrymandering in US elections
Yushu Li	Estimating APGARCH-Skew-t model by wavelet support vector machines

George A. Morcerf	The impact of the international commodity market on the Brazilian economy: an analysis using global-VAR (GVAR)
Riccardo Parviero	A viral approach to early prediction of adoptions of new products

CS-19 Inference for high-dimensional data

Contributed session, Seminari B (Seminari B)
Chair: Zbigniew Szkutnik

Stanislav Anatolyev	A ridge to homogeneity
Leonid Iosipoi	Variance Reduction via Empirical Variance Minimization
Peter Karlsson	An investigation of the performances of Liu estimators in a beta regression context in the presence of multicollinearity.
Danijel Kivaranovic	Asymptotic equivalence of inference post-model-selection with the lasso on randomized data and on a subset of the data
Luca Martino	Compressed Monte Carlo
Zbigniew Szkutnik	Discrepancy principle for Poisson inverse problems

16:30-17:30 Poster Session

Konrad Abramowicz	Family-wise error rate on domain subsets
Zahra Amini Farsani	Modeling the solar energies using statistical methods
Martina Amongero	Stochastic mixed models with different SAEM based algorithms
Georgios Aristotelous	Posterior predictive model assessment for epidemic models based on partially observed data
Xavier Bardina	Coinfection in a stochastic model for bacteriophage systems
Karim Benhenni	Local polynomial estimation of regression operators from functional data with correlated errors
Bożena Cegiełka	Hold your breath! Mathematical approach to the influence of smog on the respiratory system.
Yi-Ju Chen	Impact of multiple imputation strategies on the misspecified random effects in generalized linear mixed models for incomplete longitudinal data
Vivian Yi-Ju Chen	A comparison of geographically weighted regression models for count data
Nadia Dardenne	Dimensional reduction of a quality of life questionnaire: is confirmatory factor analysis a powerful tool?
Joanna Gołdyn	Extreme value analysis for PM2.5 air pollution in Poland
Tianyuan Guan	Sample size calculations in simple linear regression: problems and remedies
Niels Richard Hansen	Statistical learning of ordinary differential equation systems

Kai Hencken	Analysis of failure times of a system consisting of a mixture of two components including inspections
Sondre Hølleland	Spatio-temporal ARMA-GARCH model
Zuzana Hubnerova	Trends modelling in fire weather index monthly maxima
Dimitris Ioannidis	Local linear estimators in quantile regression with errors-in-variables
Seongjae Joo	Bootstrap of nonparametric dynamic discrete choice model
Jan Kalina	Nonparametric tests of symmetry for non-elliptical distributions
Min Jung Kim	Nonparametric regression Kriging with varying coefficient regression models
Tereza Konečná	Spatial analysis of laser-induced breakdown spectroscopy data for a sandstone sample
Dominika Korbas	Statistical analysis of the impact of air pollution on the circulatory system
Kyeongjun Lee	Approximate maximum product of spacing estimation of the parameter for a half-logistic distribution based on progressive censoring
Pai-Ling Li	Functional data clustering and missing value imputation
Kuo-Chin Lin	Modified information matrix tests for detecting misspecification in the random-effects structure of cumulative logit models
Yi-Chen Lin	The determinants of food waste: evidence from a semiparametric functional-coefficient cointegration test
Marco Longfils	Single diusing particles observed through a confocal microscope: an application of the doubly stochastic Poisson point process.
Benjamin Marshall	Bayesian semi-parametric analysis of multivariate continuous responses with variable selection
Vishal Midya	Bias in Bayes factor and calibrated Bayes factor for interval null hypothesis
Wessel H. Moolman	Applications of the negative hypergeometric distribution
Thi Kim Hue Nguyen	Guided structure learning of DAGs for count data
Thobeka Nombebe	Comparing the impact of using restricted against unrestricted residuals in bootstrap-based hypothesis testing in a simple regression model
Ivan Papić	Parameter estimation for non-stationary Fisher-Snedecor diffusion
Thi Huong Phan	Hierarchical spatial survival models with application to life-cell imaging data
Martina Raggi	Exact parametric causal mediation analysis for binary outcomes with binary mediators
Denise Rava	Causal Quantile Learner: causal inference for structural equation model
Jose Rodríguez-Avi	An economic application of the CTP distribution
Carles Rovira	Strong approximations of Brownian sheet by uniform transport processe
Sara Salvador	Bayesian test of bimodality for the generalized Von Miss distribution
Istoni Sant'Ana	A Bayesian non-homogeneous Markov chains to modeling and analyzing multiple sclerosis progression
Giovanni Saraceno	Robust multivariate estimation based on statistical depth filters
Martin Schindler	A new multivariate two-sample rank test

Rowland Seymour	Bayesian non-parametric inference for stochastic epidemic models
Miroslav Šiman	New advances in multiple-output quantile regression
Neill Smit	A comparison of two Bayesian accelerated life testing models
Stefan Stein	A Sparse Beta model with covariates
Nenad Suvak	Statistical analysis of stationary Fisher-Snedecor diffusion
Patrik Tardivel	On the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising
Olympia Tumolva	Model-based approach in dissolution profile comparison: alternative to f2 metric
Jan Van Waaij	Detection of communities in the stochastic block model: consistency and confidence sets
Kata Vuk	Change-point detection based on weighted two-sample U-statistics
Zhenggang Wang	CLT for linear spectral statistics of a class of Wigner-type matrices with general variance profiles
Cuiling Wang	Evaluate association and diagnosis of disease using longitudinal markers with application to Alzheimer's Disease neuropathology
Ke Yu	Depth for Riemannian manifold-valued functional data based on optimal transport

17:30-18:30 Plenary Session

PL-03 BS-EMS Lecture

Plenary, Aula Magna - Building 13
Chair: Mats Gyllenberg

Aad van der Vaart	Nonparametric Bayes: review and challenges

Wednesday, July 24

09:00-10:00 Plenary Session

PL-04 Forum Lecture I

Plenary, Aula Magna - Building 13
Chair: Irène Gijbels

Victor Panaretos	Amplitude and phase variation of random processes: part I

10:30-12:30 Invited and Contributed Sessions

IS-05 Bayesian nonparametrics and uncertainty quantification

Invited session, Room 12 (Aula 12)
Organizer: Ismael Castillo
Chair: Ismael Castillo

Eduard Belitser	Robust inference for general framework of projection structures
Botond Szabo	Spike and slab priors: recovery, uncertainty quantification and computational issues
Keisuke Yano	On frequentist coverage errors of Bayesian credible sets in moderately high dimensions

IS-13 Directional statistics

Invited session, Room 11 (Aula 11)
Organizer: Christophe Ley
Chair: Christophe Ley

Rosa M. Crujeiras	On smoothing methods for circular regression with categorical data
Peter Jupp	Parallel lines and rotations of crystals
Davy Paindaveine	Inference for spherical location under high concentration

IS-14 Statistical analysis of non-standard data types

Invited session, Room 9 (Aula 9)
Organizer: Regina Liu
Chair: Regina Liu

George Michailidis	Change point estimation in a dynamic stochastic block model
Peter X. K. Song	Regression analysis of networked data
Shahin Tavakoli	A Spatial modeling approach for linguistic object data: analysing dialect sound variations across Great Britain

TCS-14 Piecewise deterministic Monte Carlo methods

Topic-contributed session, Room 7 (Aula 7)
Organizer: Kengo Kamatani
Chair: Kengo Kamatani

Christophe Andrieu	Hypocoercivity of some PDMP Monte Carlo
Joris Bierkens	Spectral theory of the zigzag sampler
George Deligiannidis	The bouncy particle sampler and randomized HMC
Alain Durmus	On PDMP and their invariant measure
Chris Sherlock	Hug and Hop: explicit, non-reversible, contour-hugging MCMC
Kengo Kamatani	Scaling limits of piecewise deterministic Monte Carlo methods

TCS-15 Recent developments in statistical inference for stochastic processes

Topic-contributed session, Room 5 (Aula 5)
Organizer: Hiroki Masuda
Chair: Yasutaka Shimizu

Alessandro De Gregorio	Regularization methods for stochastic differential equations
Teppei Ogihara	Parameter estimation for misspecified diffusion with market microstructure noise
Yasutaka Shimizu	Model selection for determinantal point processes
Masayuki Uchida	Parametric inference for a discretely observed SPDE

CS-20 Recent advances in stochastic modeling

Contributed session, Room 10 (Aula 10)
Chair: Zhixiang Zhang

Markus Bibinger	Parameter estimation for stochastic PDEs using high-frequency observations
Claudia Furlan	How simultaneous confidence regions act in nonlinear diffusion models
Johannes Heiny	High-dimensional Kendall's Tau and Spearman's Rho correlation matrices
D. Márquez-Carreras	Semilinear fractional stochastic differential equations driven by a γ-Hölder continuous signal with $\gamma > 2/3$.
Nestor Parolya	Tests on the block-diagonal covariance matrix with a large number of blocks
Zhixiang Zhang	Asymptotic independence of spiked eigenvalues and linear spectral statistics for large sample covariance matrices

CS-21 Stochastic processes

Contributed session, Room 8 (Aula 8)
Chair: Nuno Picado

Paolo E. Cardone	Public support for an EU-wide social benefit scheme: evidence from Round 8 of the European Social Survey (ESS)
Martina Favero	A dual process for the coupled Wright-Fisher diffusion
Chrysoula Ganatsiou	On the expected extinction time for the adjoint birth-death circuit chains in random environments
Emanuele Gramuglia	A fault prediction and classification method for temporal data
Pierre-Yves Louis	Is-ClusterMPP: clustering algorithm through point processes and influence space towards high-dimensional data
Nuno Picado	Deciding about the emptiness of the interior of a manifold based on a dependent sample of points

CS-22 New developments in multivariate analysis

Contributed session, Room 6 (Aula 6)
Chair: Helena S. Rodrigues

Farrukh Javed	Fourth cumulant for multivariate aggregate claim models
Artür Manukyan	Graph-based parameter-free clustering algorithms based on Ripley's K-function
Chiara Masci	Evaluating class effects on the joint student achievements in different subjects: a bivariate semi-parametric mixed-effects model

Carel Peeters	Markov properties of the common factor analytic model
Constantinos Petropoulos	Componentwise estimation of ordered scale parameters of two exponential distributions under a general class of loss function
Helena S. Rodrigues	The influence of sponsorship in the intention purchase of Portuguese consumers

CS-23 Likelihood based modeling

Contributed session, Seminari B
Chair: Paolo Vidoni

Maryam Alamil	A statistical learning approach to infer transmissions of infectious diseases from deep sequencing data
Alessandro Barbiero	Generating correlated discrete data through the t copula
Valentina Cueva López	A new regression model
Estelle Kuhn	Estimation of the Fisher information matrix in latent variables models based on the score function
Rosalba Radice	Recursive copula additive models to estimate the effect of a binary endogenous variable in a count regression
Paolo Vidoni	Inference for multiplicative model combination using score matching

Thursday, July 25

09:00-10:00 Plenary Session

PL-05 Forum Lecture II

Plenary, Aula Magna - Building 13
Chair: Lutz Dümbgen

Victor Panaretos	Amplitude and phase variation of random processes: part II

10:30-12:30 Invited and Contributed Sessions

IS-09 Data analytics

Invited session, Room 12 (Aula 12)
Organizer: Yannig Goude
Chair: Yannig Goude

Matteo Fasiolo	Non-parametric probabilistic regression modelling of electricity demand data
Davide Ferrari	Ranking the importance of genetic factors by variable-selection confidence sets
Pierre Gaillard	Target tracking for contextual bandits: application to demand side managemen

IS-12 Statistical analysis of network data

Invited session, Room 11 (Aula 11)
Organizer: Eric Kolaczyk
Chair: Pierre Barbillon

Mingli Chen	Nonlinear factor models for network and panel data
Marianna Pensky	Estimation and clustering in popularity adjusted stochastic block model
Timothée Tabouy	Variational inference for stochastic block models from sampled data

IS-16 Dependence modelling with copulas

Invited session, Room 9 (Aula 9)
Organizer: Johanna Neslehova
Chair: Johanna Neslehova

Irène Gijbels	Test for covariate effects in conditional copula models
Thibault Vatter	Solving estimating equations with copulas
Stanislav Volgushev	Estimating the extremal dependence structure of time series extremes using block maxima

TCS-16 Advanced statistical modeling in particle physics and astronomy

Topic-contributed session, Room 7 (Aula 7)
Organizer: Alessandra R. Brazzale
Chair: Mauro Bernardi

Alessandro Casa	Nonparametric semisupervised classification and variable selection for new physics searches
Denise Costantin	A novel approach for pre-filtering event sources using the Von Mises distributions
Alex Geringer-Sameth	Dark matter interpretations of gamma-ray signals in the presence of unknown astrophysical backgrounds
Andrea Sottosanti	Astronomical source detection and background separation via hierarchical Bayesian nonparametric mixtures
Pietro Vischia	The DAB: Detecting Anomalies via Bootstrapping

TCS-17 YUIMA package: high-frequency and computational statistics

Topic-contributed session, Room 5 (Aula 5)
Organizer: Kengo Kamatani
Chair: Kengo Kamatani

Shoichi Eguchi	Stepwise model selection for SDEs in YUIMA
Emanuele Guidotti	Towards coding of the asymptotic expansion formula in YUIMA
Yuta Koike	High-dimensional covariance estimation in YUIMA package
Lorenzo Mercuri	Finite mixture approximation of CARMA(p,q) model
Yuma Uehara	Noise estimation for ergodic Levy driven SDE in YUIMA package

CS-24 Graph theory and limit theorem for stochastic processes

Contributed session, Room 10 (Aula 10)
Chair: Fanny Villers

Istvan Fazekas	Asymptotic properties of a random graph evolution model driven by a branching process
Raúl Hernández-Molinar	On the application of empirical modelling using extreme values theory to propose acceptance limits in quality control processes
Adam Jakubowski	A new central limit theorem for GARCH processes without Kesten's regularity
Ivo Stoepker	Parameter estimation of finite mixtures based on the empirical identity process
Fanny Villers	Graph inference with clustering and false discovery rate control

CS-25 Markov models and applications

Contributed session, Room 8 (Aula 8)
Chair: Nadarajah Ramesh

Leila Hosseini	Optimal investment for an insurer under Levy process with MVC criterion
Thomas Hotz	Analysing Markov chains using random measures
Barbara Jasiulis-Gołdyn	Extremal Markov chains driven by the Kendall convolution
Alastair Lamont	A missing value approach for breeding value estimation
Emilia Pompe	A framework for adaptive MCMC targeting multimodal distributions
Nadarajah Ramesh	Doubly stochastic exponential pulse models for rainfall

CS-26 Advanced topic in regression analysis and change-point detection

Contributed session, Room 6 (Aula 6)
Chair: Martin Tveten

Prajamitra Bhuyan	Analysing causal effect of London cycle superhighways on traffic congestion
Anja Mühlemann	Optimal solutions to the isotonic regression problem
Ajmal Oodally	Convergent estimation algorithm for frailty models based on integrated partial likelihood
Markus Pohlmann	Bump detection in the presence of dependency
Georgy Sofronov	Change-point modelling with applications in early detection of students at risk

Martin Tveten	Tailoring PCA for detecting sparse changes in multi-stream data

CS-27 Recent advances in robust statistics

Contributed session, Seminari B
Chair: Duc-Khanh To

Amenah Abdulateef	Weighted least squares estimators for the Parzen tail index
Hameed Al-Najafi	
Melisande Albert	Separation rates for independence tests based on wavelet decomposition
Irène Gannaz	Inference of dependence graphs by multiple testing, with application to brain connectivity
Halehsadat Nekoee Zahraei	Missing value imputation in cluster analysis
Christof Schötz	Convergence rates for the (generalized) Fréchet mean via the quadruple inequality
Duc-Khanh To	A mean score equation-based approach to correct for nonignorable verification bias in estimation of the volume under the ROC surface

14:00-16:00 Invited and Contributed Sessions

IS-01 Data privacy and protection

Invited session, Room 12 (Aula 12)
Organizer: Louis Aslett
Chair: Louis Aslett

Louis Aslett	Towards encrypted inference for arbitrary models
Murray Pollock	Confusion: developing an information-theoretic secure approach for multiple parties to pool and unify statistical data, distributions and inferences.
Sinan Yıldırım	Exact MCMC with differentially private moves

IS-21 Partially observed functional data

Invited session, Room 11 (Aula 11)
Organizer: Laura Sangalli
Chair: Alessia Pini

Antonio Elías	Depth measures for partially observed functional data
David Kraus	Functional data analysis and censoring
Dominik Liebl	Reconstructing partially observed functional data with (non-)systematically missing parts

IS-23 Stein's method and information theory

Invited session, Room 9 (Aula 9)
Organizer: Yvik Swan
Chair: Yvik Swan

Benjamin Arras	On Stein's method for multivariate self-decomposable laws with finite first moment
Yvik Swan	Stein kernel representations and their applications
Ivan Nourdin	Berry-Esseen bounds in the Breuer-Major CLT and Gebelein's inequality

TCS-18 Recent advances in time series analysis and space-time models

Topic-contributed session, Room 7 (Aula 7)
Organizer: Ta-Hsin Li
Chair: Ta-Hsin Li

Bei Chen	Recurrent neural networks for ARMA model selection
Liudas Giraitis	Robust tests for white noise and cross-correlation
Steffen Grønneberg	On partial-sum processes of ARMAX residuals
Yiannis Kamarianakis	How many filters are enough? A new algorithm for eigenvector space-time filtering
Ta-Hsin Li	Quantile-frequency analysis and functional principal components for time series classification

TCS-19 Recent advances in simulation-based methods for numerical integration and inference

Topic-contributed session, Room 5 (Aula 5)
Organizer: Luca Martino
Chair: Luca Martino

Petar Djuric	Inferring causality by Gaussian processes
Lorenzo Frattarolo	Stirring the mix: negative association and permutation polytopes
Javier Lopez-Santiago	Bayesian inference for parameter estimation in high dimension space in Astrophysics
Joaquin Miguez	Nested filters for joint parameter estimation and tracking of state-space nonlinear systems
Jesse Read	Multi-output chain models and their application in data streams
Francisco Ruiz	A contrastive divergence for combining variational inference and MCMC

CS-28 Extreme value theory and probability inequalities

Contributed session, Room 10 (Aula 10)
Chair: Edoardo Vignotto

Pascal Dkengne Sielenou	Estimation of the extrapolation range associated with extreme-value models: Application to the assessment of sensors reliability
Kaushik Jana	Scoring predictions at extreme quantiles
Gilles Stupfler	Extremiles: a new perspective on asymmetric least squares
Edoardo Vignotto	Extreme value theory for open set classification - GPD and GEV classifiers

CS-29 Recent advances in asymptotic and queueing theory

Contributed session, Room 8 (Aula 8)
Chair: Alberto Pessia

Lutz Duembgen	Bounding distributional errors via density ratios
Lothar Heinrich	Large domain statistics for Brillinger-Mixing spatial point processest
Alisa Kirichenko	Revisiting minimax property
Sofiane Ouazine	Uncertainty analysis of the GI/M/1 queue with negative customers
Alberto Pessia	Numerical evaluation of the transition probability of the simple birth-and-death process

CS-30 Recent advances in decision and information theory

Contributed session, Room 6 (Aula 6)
Chair: Jonas Moss

Dragi Anevski	Estimation of discrite monotone distribution with a prior model selection
Stefan Heyder	Non-asymptotic, universal confidence sets for intrinsic means on the circle by mass concentration
El Mehdi Issouani	Automatic text simplification
Gytis Kulaitis	What is the resolution of a microscope? A statistical minimax point of view
Bo Li	Design and implementation of recommender based on debiased learning to rank
Jonas Moss	R squared and decision theory

CS-31 Recent advances in supervised learning

Contributed session, Seminari B
Chair: Sung-Chiang Lin

Christopher Dunderdale	A machine learning approach to fault classification in photovoltaic systems using infrared imagery
Riccardo Giubilei	ETrees: a generalization of conditional trees to mixed-type data
Amine Ounajim	Mixture of varying coefficient models with random effects
Dragana Radojicic	Machine learning in finance
Oleg Sysoev	PSICA: decision trees for probabilistic subgroup identification with categorical treatments
Chun Yi Yeh	Develop a risk prediction model for depression based on text/documents mining techniques

16:30-17:30 Plenary Session

PL-06 Special Invited Lecture

Plenary, Aula Magna - Building 13
Chair: Ingrid Glad

Gilles Blanchard	Sketched learning using random moments

Friday, July 26

09:00-11:00 Invited and Contributed Sessions

IS-07 Multiscale methods

Invited session, Room 12 (Aula 12)
Organizer: Idris Eckley
Chair: Idris Eckley

Kathryn Leeming	Local white noise testing with wavelets
Rainer Von Sachs	Intrinsic wavelet regression for curves and surfaces of Hermitian positive definite matrices
Idris Eckley	A wavelet-based imputation approach for multivariate time series

IS-08 Recent advances in statistical inference with big data

Invited session, Room 11 (Aula 11)
Organizer: Yingying Fan
Chair: Yingying Fan

Thomas Berrett	The conditional permutation test for independence while controlling for confounders
Marina Bogomolov	Controlling FDR while highlighting selected discoveries
Yuekai Sun	Valid inference in bandit problems

IS-15 Inference in statistical inverse problems

Invited session, Room 9 (Aula 9)
Organizer: Axel Munk
Chair: Lutz Dümbgen

Katharina Proksch	Tests for qualitative features in the random coefficients model
Kolyan Ray	Nonparametric statistical inference for the drift of a multidimensional diffusion
Frank Werner	Statistical inference for molecules: how many and where?

TCS-20 Distance correlation and kernel independence tests

Topic-contributed session, Room 10 (Aula 10)
Organizer: Dominic Edelmann
Chair: Dominic Edelmann

Dominic Edelmann	A simple yet rigorous introduction into distance correlation
Tamara Fernandez	RKHS testing for independence under right-censored data
Jochen Fiedler	The dcortools R package: fast algorithms for the calculation of the distance correlation and applications
Mahsa Ghanbari	The distance precision matrix: computing networks from non-linear relationships
Dino Sejdinovic	RKHS mean embeddings and hypothesis testing

TCS-05 Stochastic orders and applications

Topic-contributed session, Seminari B
Organizer: Milto Hadjikyriakou
Chair: Milto Hadjikyriakou

Idir Arab	Some recent results on generalized ageing orderings with applications
Félix Belzunce	Testing stochastic dominance for dependent random variables (paired data)
Tommaso Lando	Generalizing stochastic orders through probability transformation functions
Rosa Lillo	Extremality order and extremes detection in the Cantabrian coast
Tilo Wiklund	Probabilities of exceeding mean or mode: ordering the Beta-distributions by skewness

CS-32 Advanced topic in non-standard modeling

Contributed session, Room 5 (Aula 5)
Chair: Chi Zhang

Anne Gegout Petit	Network inference for truncated gaussian data
Wen-Han Hwang	Right-censored mixed Poisson count models with detection times
Saebom Jeon	A multidimensional latent variable analysis for longitudinal data
Matthew Schofield	Latent class models for diagnostic testing with no gold standard
Chi Zhang	Valid properties of truncated Student-t distribution with applications in the analysis of censored data

CS-33 Recent advances in semiparametric modeling

Contributed session, Room 6 (Aula 6)
Chair: Motahareh Parsa

Hilal Arslan	The impacts of PM10 and SO2 concentrations on hospital admissions for pneumonia and chronic obstructive pulmonary disease in canakkale, Turkey
Gioia Di Credico	A Bayesian approach to estimate the number and position of knots for linear regression splines
Jeongmin Jeon	Additive regression with mixed predictors and incomplete Hilbertian responses
Motahareh Parsa	On AFT mixture cure models, benefits and estimation

CS-34 New developments in multivariate models

Contributed session, Room 7 (Aula 7)
Chair: Inês Sousa

Melih Agraz	Active learning and experimental design
Aoibheann Brady	A systematic checklist for causal assessment of environmental observational studies
Manuela Cattelan	Plackett models for dependent discrete data
Mauro Gasparini	Proper likelihood ratio based ROC curves
Wei Hsiang Lin	Build imbalanced classifier via kernel trick and evaluate its effect
Inês Sousa	Longitudinal models with informative time measurements

CS-35 New methods in multivariate analysis

Contributed session, Room 8 (Aula 8)
Chair: Marco Stefanucci

Hojjatollah Farahani	Fuzzy item ambiguity analysis in psychological testing and measurement
Olga Gorskikh	changedetection: an R package for nonparametric structural change detection in multivariate systems
Vojtěch Kika	Multivariate associations measures
Xiaoran Lai	Likelihood-free inference of multi-scale mathematical model for personalised breast cancer treatment
Marco Stefanucci	Overlap group Lasso in functional regression

11:30-12:30 Plenary Session

PL-07 Closing Lecture

Plenary, Aula Magna - Building 13
Chair: Anne Gegout-Petit

John Lafferty	Computational perspectives on some statistical problems

Abstracts

Estimating the reach of a manifold

Tuesday 23
10:30 - 12:30
Room 12

Eddie Aamari[1]

[1]Université Paris Diderot, Paris, France
eaamari@ucsd.edu

Various problems within computational geometry and manifold learning encode geometric regularity through the so-called reach, a generalized convexity parameter. The reach τ_M of a submanifold $M \subset \mathbb{R}^D$ is the maximal offset radius on which the projection onto M is well defined. The quantity τ_M renders a certain minimal scale of M, giving bounds on both maximum curvature and possible bottleneck structures. In this talk, we will study the geometry of the reach through an approximation theory perspective. We present new geometric results on the reach of submanifolds without boundary. An estimator $\hat{\tau}_n$ of τ_M will be described, in an idealized i.i.d. framework where tangent spaces are known. The estimator $\hat{\tau}_n$ is showen to achieve uniform expected loss bounds over a $\mathcal{C}^3$-like model. Minimax upper and lower bounds are derived. We will conclude with an extension to a model in which tangent spaces are unknown.

Probabilistic Networks for GPS Data Analysis

Monday 22
10:30 - 12:30
Room 6

Antonino Abbruzzo[1] and Angelo Mineo and Stefano De Cantis and Mauro Ferrante

[1]Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
antonino.abbruzzo@unipa.it

Nowadays, GPS devices have become of small size, not bulky, equipped with significant autonomy and, what matters most, once activated manage to memorise the geographical coordinates in which a statistical unit is at a given moment. GPS devices can record units' travel times and coordinates of locations every second. Once the experience is over, it is possible to download the data and analyse the trajectories taken by the units. These devices have several advantages and give the opportunity to collect high-quality data, that are very accurate both in terms of temporal (seconds) and spatial (meters) resolution. Indeed, these devices produce a large number of trajectories, so there is an increasing need for statistical methods that allow extracting knowledge from this data. In this paper, we propose to GPS

data statistical techniques i) for the identification of the so-called stops, and ii) derive a probabilistic network of stops that synthesized the movements of the units in a conditional independence graph. In particular, a density cluster-based algorithm is used to summarise the vast amount of information collected by GPS devices. This first step is important to determine the stops which can be regarded as an important part of a trajectory that is relevant for an application and it is characterised by consecutive GPS points which are within a predefined time and distance threshold. These stops are analyzed through a probabilistic network analysis approach in order to estimate the conditional probabilities. We apply the proposed approach to two cases study on cruise passengers' movements in two urban contexts.

Thursday 25
10:30 - 12:30
Seminari B

Weighted Least Squares Estimators for the Parzen Tail Index

Amenah Abdulateef Hameed Al-Najafi[1] and László Viharos

[1]Mathematics department, Szeged University, Szeged, Hungarian
ma_rw114@yahoo.com

Estimation of the tail index of heavy-tailed distributions and its applications are essential in many research areas. Holan and McElory [3] introduced a regression estimator for the Parzen tail index based on ordinary least squares. We propose a class of weighted least squares (WLS) estimators for the Parzen tail index. We investigate consistency and asymptotic normality of the WLS estimators. Through a simulation study, we make a comparison with the Hill [2], Pickands [4], and DEdH (Dekkers, Einmahl and de Haan) [1] estimators using the mean square error as criterion. The results show that some members of the WLS estimators are either competitive with or superior to the other estimators investigated. References [1] Arnold LM Dekkers, John HJ Einmahl, and Laurens De Haan. A moment estimator for the index of an extreme-value distribution. The Annals of Statistics, pages 1833–1855, 1989. [2] Bruce M Hill. A simple general approach to inference about the tail of a distribution. The annals of statistics, pages 1163–1174, 1975. [3] Scott H Holan and Tucker S McElroy. Tail exponent estimation via broadband log density-quantile regression. Journal of Statistical Planning and Inference, 140(12):3693–3708, 2010. [4] James Pickands III et al. Statistical inference using extreme order statistics. the Annals of Statistics, 3(1):119–131, 1975.

Tuesday 23
16:30 - 17:30
Second Floor

Family-Wise Error Rate on Domain Subsets

Konrad Abramowicz[1] and Alessia Pini and Lina Schelin and Simone Vantini

[1]Department of Mathematics and Mathematical Statistics, Umea University, Umea, Sweden
konrad.abramowicz@umu.se

A functional test of the hypotheses H0 against H1 (e.g., a test on parameters of a functional-on-scalar linear model) is considered where the aim is to select the parts of the domain where H0 is violated, while controlling the probability of false discoveries. It is straightforward

to define an unadjusted p-value function, associating a p-value with each point of the domain. Such a function only point-wise controls the probability of type I error, so it cannot be used for domain-selection purposes, since it would not provide any global control of the probability of type-I error. That is why the definition of an adjusted p-value function provided with a stronger control is often required. We require the control of the probability of falsely rejecting the null hypothesis on subsets of the domain (control of the family-wise error rate, FWER on subsets). We compare different methods to define the adjusted p-value function. The methods that we discuss belong to a general set of methods based on the following steps: a family S of subsets of the domain is defined; the restriction of the null hypothesis is tested on every element of the family; the adjusted p-value of each point is computed as the maximum p-value of the tests of every element containing that point. We consider several methods where the choice of S is either fixed or data-driven.

Flexible Modelling and Expectile Regression

Monday 22
14:00 - 16:00
Room 8

Cécile Adam[1] and Irène Gijbels

[1]Department of Mathematics, KU Leuven, Leuven, Belgium
cecile.adam@kuleuven.be

Among the main interests in regression analysis is to explore the influence that covariates have on a variable of interest, the response. There is an extensive literature on flexible mean regression, in which the targeted quantity is the conditional mean of the response given the covariates. Quantile regression is another method which aims at estimating the conditional median or other quantiles of the response variable given the covariates. An alternative to quantiles are expectiles. Expectile regression estimates the conditional expectiles of the response variable given certain values of the predictor variables. After a brief introduction to expectiles we consider nonparametric estimation of regression expectiles by using an asymmetric least squares approach. Local polynomial fitting will be used to compute the estimates of the nonparametric regression. Further, we will discuss some details for choosing the bandwidth of the nonparametric expectile regression estimator. Two bandwidth selection procedures will be discussed and compared via simulations for the local linear fitting case.

Weighted Local Second-Order Statistics for Complex Spatio-Temporal Point Processes

Monday 22
10:30 - 12:30
Room 6

Giada Adelfio[1] and Marianna Siino and Jorge Mateu and Francisco J. Rodríguez-Cortés

[1]Dipartimento di Scienze Economiche Aziendali e Statistiche, Univeristy of Palermo, Palermo, Italy
giada.adelfio@unipa.it

Spatial, temporal, and spatio-temporal point processes, and in particular Poisson processes, are stochastic processes that are largely used to describe and model the distribution of a wealth of real phenomena. When a model is fitted to a set of random points, observed in a

given multidimensional space, diagnostic measures are necessary to assess the goodness-of-fit and to evaluate the ability of that model to describe the random point pattern behaviour. The main problem when dealing with residual analysis for point processes is to find a correct definition of residuals. Diagnostics of goodness-of-fit in the theory of point processes are often considered through the transformation of data into residuals as a result of a thinning or a rescaling procedure. We alternatively consider here second-order statistics coming from weighted measures. Motivated by Adelfio and Schoenberg (2010) for the spatial case, we consider here an extension to the spatio-temporal context in addition to focussing on local characteristics. Then, rather than using global characteristics, we introduce local tools, considering individual contributions of a global estimator as a measure of clustering. Generally, the individual contributions to a global statistic can be used to identify outlying components measuring the influence of each contribution to the global statistic. In particular, our proposed method assesses goodness-of-fit of spatio-temporal models by using local weighted second-order statistics, computed after weighting the contribution of each observed point by the inverse of the conditional intensity function that identifies the process. Weighted second-order statistics directly apply to data without assuming homogeneity nor transforming the data into residuals, eliminating thus the sampling variability due to the use of a transforming procedure. We provide some characterisations and show a number of simulation studies.

Tuesday 23
10:30 - 12:30
Seminari B

Gaussian Graphical Model Under Different Model Selection Criteria for Lasso Regression

Melih Agraz[1]

[1]Giresun University, Giresun, Turkey
melih.agraz@yahoo.com

The Gaussian graphical model (GGM) is one of the famous undirected modeling approaches, which shows the graphical interactions over a set of random variables that represents conditional independencies under the multivariate normal distribution. GGM assumes that the dataset has the multivariate normal distribution and the interactions in the systems are described by the precision matrix, Θ, i.e., $Theta = \Sigma^{-1}$. In gene networks, the number of nodes p is greater than the number of observations n, i.e., $p \gg n$, that leads to the singularity problem. There are some approaches to estimate the model parameters of GGM. One of them is lasso type regression modeling. This model construct a regression model with the remaining nodes for each node when establishing the relationship between nodes. In this study, we used lasso type regression modeling for graphical models for different model selection criteria, then we propose a new model selection criteria based on the information complexity (ICOMP) and its close alternatives, namely, CAIC (consistent AIC) and CAICF (consistent AIC with Fisher Information Matrix). We compared the results from these different model selection algorithms and the results are obtained from simulation and real data set from different model selection algorithms.

Active Learning and Experimental Design

Friday 26
09:00 - 11:00
Room 7

Melih Agraz[1] and Vilda Purutçuoğlu

[1]Giresun University, Giresun, Turkey
melih.agraz@yahoo.com

The construction of biological networks can be done via different approaches from stochastic modeling like diffusion model to deterministic approches like Boolean networks or ordinary differential equations. Among these alternatives, the most common method is the models which enable us to capture the steady-state behaviour of biological systems. Under this branch, the Gaussian graphical model is one of the leading methods due to its simplicity coming from the normality assumption in modeling and its friendly usage as it is supported by R programming language. From our previous studies, we have extended this model under long-tailed symmetric (LTS) family of distribution since this covers a wide range of densities from normal, student-t to Cauchy. Whereas, we have also seen that the performance of models is highly dependent on the model selection criteria which choose the optimal estimated model among its compatitives. Hereby, in this study, we propose new model selection criteria based on the information complexity (ICOMP) and its close alternatives, namely, CAIC (consistent AIC) and CAICF (consistent AIC with Fisher Information Matrix). These criteria are evaluated for mainly LTS graphical models by using real protein-protein interaction networks. K-means and fuzzy k- means clustering methods are applied to this dataset, then we reach the core components of the dataset and then we implement model selection methods for each component.and the performance of these listed criteria with other criterion such as RIC and STARS are assessed by various accuarcy mueasures like F-measure and precision. From the results, we observe that the proposed criteria is promising to select the true estimated model and the new information gathered from these estimated networks enables us to better understand the actual biological systems.

A Statistical Learning Approach to Infer Transmissions of Infectious Diseases from Deep Sequencing Data

Wednesday 24
10:30 - 12:30
Seminari B

Maryam Alamil[1] and Samuel Soubeyrand and Gael Thébaud

[1]INRA, Aix Marseille University, Avignon, France
maryam.alamil@inra.fr

Pathogen sequence data have been exploited to infer who infected whom, by using empirical and model-based approaches. Most of these approaches exploit one pathogen sequence per infected host unit. However, data collected with deep sequencing techniques, providing a subsample of the pathogen variants at each sampling time, are expected to give more insight on epidemiological links than a single sequence per host unit. A mechanistic viewpoint to transmission and micro-evolution has generally been followed to infer epidemiological links from these data. Here, we investigate an alternative statistical learning approach for estimating epidemiological links, which consists of learning the structure of epidemiological links with a pseudo-evolutionary model and training data before inferring links for the whole data set. We designed the pseudo-evolutionary model as a semi-parametric regression function where the response variable is the set of sequences observed from a recipient host unit and the explanatory variable is the set of sequences observed from a putative source. We derived from this model a penalized pseudo-likelihood that is used for selecting who infected whom or who is closely related to whom, where the penalization is calibrated on training data. In order to assess the efficiency of the pseudo-evolutionary model and the associated inference approach for estimating epidemiological links, we applied it to simulated data generated with diverse sampling efforts, sequencing techniques, and stochastic models of viral evolution and transmission. Then, we applied it to three real epidemics: swine Influenza, Ebola and a potyvirus of wild salsify. Such an approach has the potential to be particularly valuable in the case of a risk of erroneous mechanistic assumptions and sequencing errors, it is sufficiently parsimonious to allow handling big data sets in the future, and it can be applied to very different contexts from animal, human and plant epidemiology.

Thursday 25
10:30 - 12:30
Seminari B

Separation Rates for Independence tests based on wavelet decomposition

Mélisande Albert[1]

[1]Insa de Toulouse, Université Paul Sabatier, Toulouse, France
melisande.albert@insa-toulouse.fr

This work presents an independence testing procedure in the nonparametric density framework. First, based on a wavelet thresholding method, single coefficient tests are constructed. Their corresponding critical values are obtained from a permutation approach. In particular, each of these single tests is known to be non-asymptotically of prescribed level. Then, a multiple testing procedure based on aggregation is introduced, avoiding the delicate question of the choice of the coefficient. A non-asymptotic study of the proposed procedures is performed, providing conditions on the alternative ensuring a control of the second kind error rate by a prescribed value. This leads to an upper-bound for the uniform separation rates of the aggregated test (with respect to L2-metric) over weak Besov bodies, which should be minimax. Moreover, the whole procedure is adaptive as it is entirely data-driven, and does not require any knowledge on the smoothness of the alternative.

Tuesday 23
09:00 - 10:00
Aula Magna
Building 13

Data Integration: Data-Driven Discovery from Diverse Data Sources

Genevera Allen[1]

[1]Departments of Statistics, Rice University, Houston, USA
gallen@rice.edu

Data integration, or the strategic analysis of multiple sources of data simultaneously, can often lead to discoveries that may be hidden in individual analyses of a single data source. In this talk, we present several new techniques for data integration of mixed, multi-view data where multiple sets of features, possibly each of a different domain, are measured for the same set of samples. This type of data is common in heathcare, biomedicine, national

security, multi-senor recordings, multi-modal imaging, and online advertising, among others. In this talk, we specifically highlight how mixed graphical models and new feature selection techniques for mixed, mutli-view data allow us to explore relationships amongst features from different domains. Next, we present new frameworks for integrated principal components analysis and integrated generalized convex clustering that leverage diverse data sources to discover joint patterns amongst the samples. We apply these techniques to integrative genomic studies in cancer and neurodegenerative diseases to make scientific discoveries that would not be possible from analysis of a single genomics data set.

Nonparametric Drift Estimation for Linear SPDEs from Local Measurements

Tuesday 23
14:00 - 16:00
Room 9

Randolf Altmeyer[1]

[1]Humboldt University, Berlin, Germany
altmeyrx@math.hu-berlin.de

We study the problem of nonparametric estimation in a linear second order stochastic partial differential equation (SPDE). For this a new observation scheme is introduced: The solution is observed continuously in time, but spatially localised with respect to a test function. This describes locally averaged measurements, such as heat or concentration of some substance. For the asymptotic regime with fixed time horizon and with the spatial localisation tending to zero, we provide rate-optimal estimators for the coefficients of the underlying differential operator. This is statistically remarkable: While it is well-known that consistent drift estimation for stochastic ordinary differential equations is possible for large time asymptotics, small noise levels or independent observations, here we show that shrinking spatial localisation already yields consistency. What is more, even in the nonparametric setup with spatially varying coefficients, the parametric rate is achieved and central limit theorems can be proved.

Autocovariance Varieties of Moving Average Random Fields

Tuesday 23
14:00 - 16:00
Room 5

Carlos Améndola[1] and Viet Son Pham

[1]Mathematics, TU Munich, Garching (b. München), Germany
carlos.amendola@tum.de

We study the autocovariance functions of moving average random fields over the integer lattice from an algebraic perspective. These autocovariances are parametrized polynomially by the moving average coefficients, hence tracing out algebraic varieties. We derive dimension and degree of these varieties and we use their algebraic properties to obtain statistical consequences such as identifiability of model parameters. We connect the problem of parameter estimation to the algebraic invariants known as euclidean distance degree and maximum likelihood degree. Throughout, we illustrate the results with concrete examples. In our computations we use tools from commutative algebra and numerical algebraic geometry. Joint work with Viet Son Pham.

Monday 22
10:30 - 12:30
Room 12

Optimization Algorithm for Parameter Estimation in Cell Biology

Zahra Amini Farsani[1] and Volker Schmid

[1]Statistics, LMU Munich, Munich, Germany
aminifarsani.z@lu.ac.ir

Colocalization analysis is a popular technique for the quantitative analysis of fluorescence microscopy images. The intracellular localization of proteins and their specific compartments plays a key role in the study of biological processes. Several metrics have been developed for measuring the co localization of two probes, however, they depend on a subjective thresholding of background. To this end, we propose a useful method in order to estimate the bivariate distribution function of two color channels and from that determine their co- or anti-colocalization. The proposed method is a combination of the Maximum Entropy Method (MEM) and Gaussian Copula. The proposed method is compared with MEM for bivariate probability distributions and a Gaussian mixture model. The methods are compared to simulated and real data. The results show that the new algorithm can determine co- and anti-colocalization even in high background settings. It can, therefore, be used as a robust tool for colocalization analysis.

Tuesday 23
16:30 - 17:30
Second Floor

Modeling the Solar Energies using Statistical Methods

Zahra Amini Farsani[1]

[1]Statistics, Lorestan University, Khorramabad, Iran
aminifarsani.z@lu.ac.ir

The present study thus proposes utilizing an accurate numerical-probabilistic algorithm which is the combination of the classic mathematical technique and the Maximum Entropy approach to determine an important distribution in the renewable energy systems, namely the Hyper Rayleigh Distribution which belongs to the family of Weibull distribution. It mainly used to model the wind speed and the variations of the solar irradiance level with a negligible error. The purpose of this research is to find the unique solution to an optimization problem which occurs when maximizing Shannon's entropy. To confirm the accuracy and efficiency of our algorithm, we used the long-term data for the average daily wind speed in JAPAN from the National Climatic Data Center for 12 years to examine the Rayleigh Distribution (RD). It seems that the Rayleigh distribution is more closely fitted to the data.

Stochastic Mixed Models with Different SAEM Based Algorithms

Tuesday 23
16:30 - 17:30
Second Floor

Martina Amongero[1] and Gianluca Mastrantonio and Enrico Bibbona

[1]DISMA, Politecnico di Torino, Torino, Italy
martina.amongero@studenti.polito.it

Mixed models based on ordinary differential equations are a very common tool in biology and pharmacokinetics. Their extension to stochastic differential equation models is non trivial. Some recent solutions based on a Stochastic Approximation Expectation Maximization (SAEM) algorithm have been proposed (Samson et al. 2008, 2014) and compared. For specific models (e.g. if the transition density is explicitly available), they perform very well, but it is difficult to extend them in more general situations. We discuss the available solutions, their weaknesses and propose some improvements. The validity of our proposals is tested on simulated data.

Bounds for the asymptotic distribution of the likelihood ratio

Monday 22
10:30 - 12:30
Room 5

Andreas Anastasiou[1] and Gesine Reinert

[1]Department of Statistics, The London School of Economics and Political Science, London, United Kingdom
a.anastasiou@lse.ac.uk

In this talk, we give an explicit upper bound on the distance to chi-square for the likelihood ratio statistic when the data are realisations of independent and identically distributed random elements. To our knowledge this is the first explicit bound which is available in the literature. The bound depends on the number of samples as well as on the dimension of the parameter space. We illustrate the behaviour of the bound with three examples: samples from an exponential distribution, samples from a normal distribution, and logistic regression.

A Ridge to Homogeneity

Tuesday 23
14:00 - 16:00
Seminari B

Stanislav Anatolyev[1]

[1]CERGE-EI, Prague, Czech Republic
stanislav.anatolyev@cerge-ei.cz

In some heavily parameterized econometric models, one may benefit from shrinking a subset of parameters towards a common target. We consider L2 shrinkage towards an equal parameter value that balances between unrestricted estimation (i.e., allowing full heterogeneity) and estimation under equality restriction (i.e. imposing full homogeneity). The

penalty parameter of such ridge regression estimator is tuned using one-leave-out cross-validation. The reduction in predictive mean squared error tends to increase with the dimensionality of the parameter set. We illustrate the benefit of such shrinkage with a few stylized examples. We also work out, both theoretically and empirically, a heterogenous linear panel data setup and compare several estimators and corresponding confidence intervals.

Wednesday 24
10:30 - 12:30
Room 7

Hypocoercivity of some PDMP Monte Carlo

Christophe Andrieu[1] and Alain Durmus and Nick Nueksen and Julien Roussel

[1]School of Mathematics, University of Bristol, Bristol, United Kingdom
c.andrieu@bristol.ac.uk

We review recent results concerning the scaling (with dimension of the problem) of recently proposed continuous time Monte Carlo schemes relying on piecewise deterministic Markov processes.

Thursday 25
14:00 - 16:00
Room 6

Estimation of Discrite Monotone Distribution with a Prior Model Selection

Dragi Anevski[1]

[1]Lund University, Lund, Sweden
dragi.anevski@matstat.lu.se

We propose a new method of estimating a discrete monotone probability mass function. We introduce a two-step procedure: First, we perform a consistent model selection using the Akaike-type in-formation criterion (CM AIC). Second, using the selected class of models we construct a modified Grenander estimator by grouping the parameters in the constant regions and then projecting the grouped empirical estimator onto the isotonic class. We show that the new estimator performs asymptotically better, in l_2 sense, than the regular Grenander estimator.

Some Recent Results on Generalized Ageing Orderings With Applications

Friday 26
09:00 - 11:00
Seminari B

Idir Arab[1] and Paulo Eduardo Oliveira and Milto Hadjikyriakou

[1]CMUC, Unversity of Coimbra, Coimbra, Portugal
idir.bhh@gmail.com

In this talk, we are interested in the study of both generalized failure rate monotonicity and convex transform stochastic order properties of positive random variables. Especially, we are interested in iterated distributions which equivalently represents moments of higher order of the residual lifetime. With regards to failure rate properties, we are mainly concerned about the relationships among related notions of failure rate and their hereditary properties. Moreover, we study the asymptotic behavior of iterated distributions within Gamma and Weibull families which leads to an approximation of moments of higher order of the residual lifetime. Since the random variables we are studying can be used to represent lifetimes, their probability distributions cannot be symmetric, thus the convex transform order can be used to compare random variables with respect to their skewness. Unfortunately, once we leave the exponential distribution world, the actual identification of an order relationship is computationally complex. As a remedy, we introduced a new criterium based on the analysis of the sign variation of a suitable function. We use the criterium to establish order relationships either within the Gamma, Weibull families. As another application, we derive ageing properties of parallel systems formed with components that have exponentially distributed lifetimes. Moreover, we provide a partial answer for an ordering problem that remained open for a decade.

Posterior Predictive Model Assessment for Epidemic Models Based on Partially Observed Data

Tuesday 23
16:30 - 17:30
Second Floor

Georgios Aristotelous[1] and Theodore Kypraios and Philip O'neill

[1]School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom
georgios.aristotelous@nottingham.ac.uk

During the last decade there has been a significant progress in the area of parameter estimation for stochastic epidemic models. However, less attention has been given to the issue of model adequacy and assessment. The fact that epidemic data are realized once, partially observed and not independent makes the use of simple attractive measures of model fit, such as chi-squared goodness-of-fit statistics, complicated to apply. Many researchers have used the general notion of posterior predictive checking, whereby one or more discrepancy/statistic is examined for departures from its posterior predictive distribution. Whereas posterior predictive checks has proved somewhat useful as a means of excluding models on the basis of their inability to reproduce key aspects of observed epidemics, the approaches have generally suffered from low power; either due to discarding valuable information either due to imputing a high dimensional latent process. We present two novel predictitve checks based on disease progression curves (removal curves) that utilize all the observed data and are independent of latent variables. The claim is that by acknowledging the peculiarities of the epidemic setting we can greatly increase the power of these checks. The effectiveness of the approaches is exhibited via simulations.

Thursday 25
14:00 - 16:00
Room 9

On Stein's Method for Multivariate Self-Decomposable Laws with Finite First Moment

Benjamin Arras[1] and Christian Houdré

[1]Laboratoire Paul Painlevé, Université de Lille, Villeneuve-d'Ascq, France
arrasbenjamin@gmail.com

In this talk, I will present some recent results around Stein's method in a multidimensional setting. Essentially developed for the multivariate normal distribution, Stein's method can be implemented for a general class of distributions, namely the self-decomposable laws with finite first moment. Furthermore, we will see how variational techniques recently developed for the normal distribution can be imported in this infinitely divisible framework. In particular, these make it possible to obtain explicit quantitative bounds on classical probability metrics.

These results come from works in collaboration with Christian Houdré.

Friday 26
09:00 - 11:00
Room 6

The Impacts of PM10 and SO2 Concentrations on Hospital Admissions for Pneumonia and Chronic Obstructive Pulmonary Disease in Canakkale, Turkey

Hilal Arslan[1] and Hakki Baltaci and Bulent Oktay Akkoyunlu

[1]Institute of Pure and Applied Sciences, Marmara University, Istanbul, Turkey
hilal.arslan@gedik.edu.tr

As a case-crossover study, this study investigates the impacts of PM10 and SO2 concentrations on pneumonia and chronic obstructive pulmonary disease (COPD) in the Canakkale province of Anatolian Peninsula (W Turkey). Air pollutants and hospital admissions data were taken for the period 2007-2017. The data were analyzed using Poisson regression models after controlling for the long-term trend of the air pollutants via meteorological factors (i.e. daily mean, minimum and maximum temperatures and daily mean pressure, relative humidity, wind speeds), day of the week and seasonal time trend. The analysis was conducted among people of all ages, especially focused on different sexes and different age groups including children (0-14 yrs), adults (15-64 yrs), and elderly ($>$64 yrs). Effects of the pollutants (Relative Risk, RR) from current (lag 0) to the first six days (lag 6) on respiratory diseases were determined. The relative risks (RR) were calculated by considering the increases in the inter-quartile range of the pollutants (25 μgm-3 in PM10, 17 μgm-3 in SO2). Results showed that the statistically significant ($p<0.05$) highest associations between PM10 and COPD were shown for children at lag5 with 2.1 relative risk (95% CI: 0.9-4.6). The significant influence of SO2 concentrations on COPD disease were observed for adult and female groups at lag0 with 1.03% RR (95% CI: 1.01-1.05). When compared with COPD, we were unable to demonstrate significant effects of pollutants on Pneumonia possibly due to the higher misclassification.

Towards Encrypted Inference for Arbitrary Models

Thursday 25
14:00 - 16:00
Room 12

Louis Aslett[1]

[1]Department of Mathematical Sciences, Durham University, Durham, United Kingdom
louis.aslett@durham.ac.uk

There has been substantial progress in development of statistical methods which are amenable to computation with modern cryptographic techniques, such as homomorphic encryption. This has enabled fitting and/or prediction of models in areas from classification and regression through to genome wide association studies. However, these are techniques devised to address specific models in specific settings, with the broader challenge of an approach to inference for arbitrary models and arbitrary data sets receiving less attention. This talk will discuss very recent results from ongoing work towards an approach which may allow theoretically arbitrary low dimensional models to be fitted fully encrypted, keeping the model and prior secret from data owners and vice-versa. The methodology will be illustrated with a variety of examples, together with a discussion of the ongoing direction of the work.

Stable IPW Estimation for Longitudinal Studies

Tuesday 23
14:00 - 16:00
Room 12

Vahe Avagyan[1] and Stijn Vansteelandt

[1]Biometris, Wageningen University & Research, Wageningen, Netherlands
vahe.avagyan@wur.nl

In this paper, we consider estimation of the average effect of a time-varying dichotomous exposure on an outcome using Inverse Probability Weighting (IPW) under the assumption that there is no unmeasured confounding of the exposure - outcome association at each time point. Despite the popularity of IPW for this problem, the performance of IPW estimation is often poor due to instability of the estimated weights. We develop an estimating equation-based strategy for the nuisance parameters indexing the weights at each time point, aimed at preventing highly volatile weights and ensuring the stability of IPW estimation. Our proposed approach targets the estimation of the counterfactual mean under a chosen treatment regime and therefore requires fitting a separate propensity score model at each time point, although we discuss and examine extensions to enable the fitting of Marginal Structural Models using one propensity score model across all time points. Extensive simulation studies demonstrate adequate performance of the proposed approach compared with the traditional maximum likelihood estimator and the Covariate Balancing Propensity Score estimator.

Tuesday 23
10:30 - 12:30
Seminari B

A New Steady-State Modeling Approach for Protein-Protein Interaction Networks

Ezgi Ayyıldız Demirci[1] and Vilda Purutçuoğlu

[1]Department of Statistics, Middle East Technical University, Ankara, Turkey
ayezgi@metu.edu.tr

Modeling of complex biological networks plays an important role in understanding the molecular mechanism of the systems. But, due to the high dimensional, correlated and sparse structures of protein-protein interactions' (PPI) data, it is difficult to cope with the estimation of these PPIs' networks. There are different methods to construct such structures via undirected graphs. One of the widely used graphical approaches is the Gaussian graphical model (GGM). GGM estimates the interactions within proteins by using the conditional independencies under the multivariate normality assumption of the states. On the other side, the conic multivariate adaptive regression spline (CMARS), which is the conic version of MARS, is a strong and nonparametric alternate of GGM since it can also deal with the high dimensional and correlated data. In this study, we develop a new modeling approach based on CMARS for the PPIs networks. We call it the loop-based CMARS, shortly LCMARS. LCMARS preforms an iterative calculation of CMARS by using the main and the second-order interaction effects in order to represent the interaction of each gene separately. Moreover, this model describes a mathematical expression similar to the lasso type of regression models without the strict normality assumption of states. In our analysis, we compare the performance of LCMARS with MARS that is also specifically adapted to sparse biological networks, and GGM methods under various simulation scenarios generated under distinct distributions and dimensions. Additionally, several synthetic and real biological datasets are applied to evaluate the performance of these methods in terms of different accuracy measures. From the results, it is seen that LCMARS can be a plausible alternate of GGM, particularly, when the networks are high and have non-normally distributed states.

Monday 22
10:30 - 12:30
Room 7

A Smooth Goodness-of-Fit Test for Densities

Dimitrios Bagkavos[1] and Prakash Patil and Andrew Wood

[1]Mathematics and Applied Mathematics, University of Crete, Heraklion, Greece
bagavosd2000@yahoo.gr

This research contributes a novel goodness-of-fit test which converges faster to the normal distribution and has improved rate in detecting Pitman alternatives compared to the classical test of Bickel and Rosenblatt (1973). A practically useful bandwidth selector is also developed herein, based on maximization of the test's power subject to keeping the size constant. The article concludes with numerical evidence on the performance of the test in comparison to modern goodness-of-fit tests for Pitman and Kullback–Liebler alternatives as well as the two–sample setting.

Generating Correlated Discrete Data Through the T Copula

Wednesday 24
10:30 - 12:30
Seminari B

Alessandro Barbiero[1]

[1]Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy
alessandro.barbiero@unimi.it

The t copula is a member of the family of elliptical copulas, comprising among other the well-known Gauss copula, with which it shares many properties. The t copula, however, differently from the Gauss copula, which is tail-independent, always possesses non-null lower and upper tail dependence coefficients, even when its linear correlation is zero. This feature is particularly useful when it is necessary to model a non-negligible probability for simultaneous extreme events (e.g., in financial applications, joint "big" losses). While copulas represent a natural and flexible statistical tool for modeling dependence between continuous distributions, their use is less natural in the discrete case, mainly due to theoretical reasons connected with Sklar's theorem. A practical consequence is that dependence measures that are functions only of the copula in the continuous case, such as the coefficients of lower and upper tail dependence, now are no longer margin-free. In this work, we illustrate the use of the t copula for linking discrete distributions, examining the influence of the t copula parameters (correlation and degrees of freedom) and of the marginal distributions on the correlation and other dependence measures of the resulting bivariate discrete random variable, highlighting differences with using the Gauss copula, whose utilization along with discrete margins has been already explored in several contributions of the statistical literature. Aspects related to pseudo-random generation and inference are also investigated.

Coinfection in a Stochastic Model for Bacteriophage Systems

Tuesday 23
16:30 - 17:30
Second Floor

Xavier Bardina[1] and Sílvia Cuadrado and Carles Rovira

[1]Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
Xavier.Bardina@uab.cat

A system modeling bacteriophage treatments with coinfections in a noisy context is analysed. We prove that in a small noise regime, the system converges in the long term to a bacteria-free equilibrium. Moreover, we compare the treatment with coinfection with the treatment without coinfection, showing how coinfection affects the convergence to the bacteria-free equilibrium.

Monday 22
16:30 - 18:30
Room 6

Change-Point Detection in a Markov Random Field

Michael Baron[1] and Zhongyan Liang

[1]Department of Mathematics and Statistics, American University, Washington DC, United States
baron@american.edu

We study detection and estimation of change points in an anisotropic Markov random field of events occurring on a surface. Its naturally appearing canonical parameters reflect the overall intensity and also, the direction and speed of development of events. Each parameter and its estimator have a specific meaning that describes the dynamic patterns of events and connections between neighboring events on a spatio-temporal scale. A threshold is chosen to control the rate of false alarms triggered by the change-point detector. When a significant change is detected, it may indicate a breakdown, a potential problem, or even a threat such as organized crime. Applications are shown in semiconductor industry and in threat detection.

Monday 22
10:30 - 12:30
Room 9

Kernel Matrices in the Flat Limit

Simon Barthelmé[1] and Konstantin Usevich

[1]CNRS, Gipsa-lab, Univ. Grenoble Alpes & Grenoble INP, Grenoble, France
simon.barthelme@gipsa-lab.fr

Kernel matrices are ubiquitous in statistics and machine learning. Within Bayesian statistics they occur most often as covariance matrices of Gaussian processes, in non-parametric or semi-parametric models. Most of the theoretical work on kernel methods has focused on a large-n asymptotics, characterising the behaviour of kernel matrices as the amount of data increases. Fixed-sample analysis is much more difficult outside of simple cases, such as locations on a regular grid.

In this talk I will describe a fixed-sample analysis that was first studied in the context of approximation theory by Fornberg & Driscoll (2002), called the "flat limit". In flat-limit asymptotics, the goal is to characterise kernel methods as the length-scale of the kernel function tends to infinity, so that kernels appear flat over the range of the data. In Bayesian terms, we let the covariance of the Gaussian process go to 1 over all pairs of locations. While the resulting prior becomes singular, fascinatingly, GP regression is still well-defined in the limit. In the flat limit, different types of kernels behave differently, and what matters most is the smoothness of kernel functions. The flat limit also highlights the close kinship between kernel methods and polynomial and spline regression. I will discuss implications for GP regression and Determinantal Point Processes.

Berkson's Paradox and Weighted Distributions: An application to Alzheimer's Disease

Tuesday 23
10:30 - 12:30
Room 6

Apostolos Batsidis[1] and Polychronis Economou and George Tzavelas and Panagiotis Alexopoulos

[1]Department of Mathematics, University of Ioannina, Ioannina, Greece
abatsidis@uoi.gr

Berkson's paradox occurs whenever unrelated random variables become spuriously associated in an observed sample. In 1946 Joseph Berkson first illustrated the paradox with a case–control study linking diabetes with cholecystitis among inpatients who seek care. The two diseases were found to be positive correlated even if they are independent in the population. Berkson explained this spuriously finding by noting that a hospital patient without diabetes is more likely to have cholecystitis than a member of the general population, since the patient must have had some non-diabetes (possibly cholecystitis-causing) reason to enter the hospital. Since then, many authors encourage physicians to understand Berkson's paradox in order to avoid misinterpreting data whenever counter-intuitive findings are observed. In the present talk our interest is focused on how we can extract information for the entire population based on a sample in which the Berkson's paradox is observed. In this context, it is initially recognized that Berkson's paradox is actually a selection bias problem introduced by a biased selection procedure resulting to a non-representative sample of the population intended to be analyzed and the concept of a weighted distribution is proposed for describing Berkson's paradox. Afterwards, the Approximate Bayesian Computation rejection algorithm, a likelihood-free inference method, is used to make inference for the population given a biased sample which possesses all the characteristics of Berkson's paradox. The proposed method is illustrated in a real data application for patients with dementia due to Alzheimer's disease. The application reveals characteristics of the population that are masked by the sampling procedure.

Effective Probability Distributions for Spatially Dependent Processes

Monday 22
16:30 - 18:30
Room 11

Anastassia Baxevani[1] and Dionissios Hristopulos

[1]Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus
baxevani@ucy.ac.cy

Spatially distributed physical processes can be modeled as random fields. The complex spatial dependence is then incorporated in the joint probability density function. Knowledge of the joint probability density allows predicting the field values at points where measurements are missing. The probability distribution of spatially dependent processes often exhibits significant deviations from Gaussian behavior (rainfall and earthquakes being characteristic examples). However, only a few non-Gaussian joint probability density models admit explicit expressions. In addition, spatial random field models based on Gaussian or non-Gaussian joint densities incur formidable computational costs for big datasets. We propose an "effective distribution" approach which replaces the joint probability density with a product of univariate conditional probability density functions modified by a local interaction term. The effective densities involve localized parameters that link the densities

at different locations. The prediction of the field at unmeasured locations is formulated in terms of the respective effective distribution and local constraints. We also propose a sequential simulation approach for generating multiple field realizations based on the effective distribution approach. The effective probability density model can capture non-Gaussian dependence, and it can be applied to large spatial datasets, since it does not require the storage and inversion of large covariance matrices.

Wednesday 24
10:30 - 12:30
Room 12

Robust Inference for General Framework of Projection Structures

Eduard Belitser[1] and Nurzhan Nurushev

[1]Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands
e.n.belitser@vu.nl

We develop a general framework of projection structures and study the problem of inference on the unknown parameter within this framework by using empirical Bayes and penalization methods. The main inference problem is the uncertainty quantification, but on the way we solve the estimation and posterior contraction problems as well (and a weak version of the structure recovery problem). The approach is local in that the quality of the inference procedures is measured by the local quantity, the oracle rate, which is the best trade-off between the approximation error by a projection structure and the complexity of that approximating projection structure. The approach is also robust in that the stochastic part of the general framework is assumed to satisfy only certain mild condition, the errors may be non-iid with unknown distribution. We introduce the excessive bias restriction (EBR) under which we establish the local (oracle) confidence optimality of the constructed confidence ball.

As the proposed general framework unifies a very broad class of high-dimensional models interesting and important on their own right, the obtained general results deliver a whole avenue of results (many new ones and some known in the literature) for particular models and structures as consequences, including white noise model and density estimation with smoothness structure, linear regression and dictionary learning with sparsity structures, biclustering and stochastic block models with clustering structure, covariance matrix estimation with banding and sparsity structures, and many others. Various adaptive minimax results over various scales follow also from our local results.

Point Estimation Based on Confidence Intervals

Tuesday 23
10:30 - 12:30
Room 6

Ruggero Bellio[1] and Donald A. Pierce

[1]Dipartimento di Scienze Economiche e Statistiche, Università di Udine, Udine, Italy
ruggero.bellio@uniud.it

Unbiased estimation fails to transform naturally under reparameterization, while likelihood-based confidence intervals need not suffer from this defect. This tension can be rectified by using zero-level confidence intervals as a basis for point estimation, as was proposed by Skovgaard. This also provides for median unbiased estimation. Modern likelihood asymptotics provides for confidence intervals more accurate than those based on normal approximations. The coverage probability of these improved confidence intervals is to second or third order in sample size equal to the nominal value. Beyond exponential families this theory was difficult to implement, but an approximation due to Skovgaard resolved this. Our R package likelihoodAsy implements in broad generality this modern likelihood theory, based on that approximation. Basically, the user needs only to provide R code for computing the likelihood function plus a function for generating data from the chosen model. Thus, the R package enables simply and in broad generality confidence intervals and the desired point estimation. We illustrate this for some statistical models, including generalized linear models and models involving large numbers of nuisance parameters.

Testing Stochastic Dominance for Dependent Random Variables (paired data)

Friday 26
09:00 - 11:00
Seminari B

Félix Belzunce[1] and Carolina Martínez-Riquelme

[1]Dpt. Statistics and Operations Research, University of Murcia, Espinardo (Murcia), Spain
belzunce@um.es

In this talk we consider a new stochastic dominance criteria for the comparison of two dependent random variables. Whereas the usual stochastic dominance criteria compares the marginal distributions of two random variables, there are situations where the dependence structure of the two random variables should be taken into account. This is the case when comparing random variables form paired samples. Here, we present a new stochastic dominance criteria that takes into account the dependence of the random variables and we develop a test for testing the null hypothesis that the two random variables are ordered according to the new criteria versus the alternative, where they are not ordered.

This work is supported by the Ministerio de Economía, Industria y Competitividad under grant MTM2016-79943-P (AEI/FEDER, UE).

Tuesday 23
16:30 - 17:30
Second Floor

Local Polynomial Estimation of Regression Operators from Functional Data with Correlated Errors

Karim Benhenni[1] and Ali Hajj Hassan and Yingcai Su

[1]Laboratoire Jean Kuntzmann, Université Grenoble Alpes, 38401, Saint-Martin d'Hères, France
karim.benhenni@univ-grenoble-alpes.fr

This work considers the problem of nonparametric estimation of the regression operator r in a functional regression model $Y = r(x) + \varepsilon$ with a scalar response Y, a functional explanatory variable x, and a second order stationary error process ε. We construct a local polynomial estimator of r together with its Fréchet derivatives from functional data with correlated errors. The convergence in mean squared error of the constructed estimator is studied for both short and long range dependent error processes. Simulation studies on the performance of the proposed estimator are conducted, and applications to spectrometric data and El Niño time series data are given.

Monday 22
16:30 - 18:30
Room 7

CARMA Generalized Processes and Stochastic Partial Differential Equations

David Berger[1]

[1]Ulm University, Ulm, Germany
david.berger@uni-ulm.de

We give a new definition of a Lévy driven CARMA random field, defining it as a generalized solution of a stochastic partial differential equation (SPDE). Furthermore, we give sufficient conditions for the existence of a mild solution of our SPDE. Our model finds a connection between all known definitions of CARMA random fields, and especially for dimension 1 we obtain the classical CARMA process.

Friday 26
09:00 - 11:00
Room 11

The Conditional Permutation Test for Independence while Controlling for Confounders

Thomas Berrett[1] and Yi Wang and Rina Foygel Barber and Richard Samworth

[1]Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom
t.berrett@statslab.cam.ac.uk

In this talk I will discuss the discuss the problem of testing conditional independence. In standard independence testing problems, a simple and practical approach to find critical values and calibrate our tests is to uniformly at random permute the data in order to simulate the behaviour under the null hypothesis of independence. Unfortunately, this is

no longer effective when testing conditional independence, and may result in misleading conclusions. We propose a general new method, the conditional permutation test, for testing the conditional independence of variables X and Y given a potentially high-dimensional random vector Z that may contain confounding factors. The proposed test permutes entries of X non-uniformly, so as to respect the existing dependence between X and Z and thus account for the presence of these confounders. Like the conditional randomization test of Candès et al., our test relies on the availability of an approximation to the distribution of X|Z—while Candès et al.'s test uses this estimate to draw new X values, for our test we use this approximation to design an appropriate non-uniform distribution on permutations of the X values already seen in the true data. We provide an efficient Markov Chain Monte Carlo sampler for the implementation of our method, and establish bounds on the Type I error in terms of the error in the approximation of the conditional distribution of X|Z, finding that, for the worst case test statistic, the inflation in Type I error of the conditional permutation test is no larger than that of the conditional randomization test. We validate these theoretical results with experiments on simulated data and on the Capital Bikeshare data set.

Optimal Transport Methods in Statistics and Machine Learning

Tuesday 23
14:00 - 16:00
Room 11

Quentin Berthet[1] and Jonathan Weed

[1]DPMMS Statslab, University of Cambridge, Cambridge, United Kingdom
q.berthet@statslab.cam.ac.uk

Optimal transport is one of the foundational problems of optimization, and a very important topic in analysis. It asks how one can transport mass with a given measure to have another measure, with minimal global transport cost. The associated Wasserstein distance is a useful tool to compare distributions, taking into account geometric properties of the data. In this talk, I will present some recent results in this area.

Testing for a Change in the Tail Index of Long-Memory Stochastic Volatility Time Series

Monday 22
16:30 - 18:30
Room 8

Annika Betken[1] and Rafał Kulik and Davide Giraudo

[1]Ruhr-Universität Bochum, Bochum, Germany
annika.betken@rub.de

Let X_k, $k \in \mathbb{Z}$, be a stationary time series with tail index $\alpha > 0$, i.e. $\bar{F}(x) := P(X_1 > x) = x^{-\alpha}L(x)$, where L is a slowly varying function. Then, we have

$$\lim_{u\to\infty} \mathrm{E}\left(\log\left(\frac{X_1}{u}\right) \mid X_1 > u\right) = \frac{1}{\alpha} =: \gamma,$$

such that γ can be estimated by the Hill estimator

$$\hat{\gamma} := \frac{1}{\sum_{j=1}^{n} 1_{\{X_j > u_n\}}} \sum_{j=1}^{n} \log\left(\frac{X_j}{u_n}\right) 1_{\{X_j > u_n\}} ,$$

where u_n, $n \geq 1$, is a deterministic sequence with $u_n \to \infty$ and $n\bar{F}(u_n) \to \infty$. We consider a change-point test based on the Hill estimator to test for structural changes in the tail index of long-memory stochastic volatility (LMSV) time series. In order to determine the asymptotic distribution of the corresponding test statistic, we prove a uniform reduction principle for the tail empirical process in a two-parameter Skorohod space. It is shown that such a process displays a dichotomous behavior according to an interplay between the Hurst parameter, i.e. a parameter characterizing the dependence in the data, and the tail index. We will see that, nonetheless, long-memory does not have an influence on the asymptotic behavior of the test statistic.

Monday 22
16:30 - 18:30
Room 5

Invariant Aspects of Functional Data with Phase Variation

Karthik Bharath[1] and Ian Jermyn

[1], University of Nottingham, Nottingham, United Kingdom
Karthik.Bharath@nottingham.ac.uk

Phase variation is a unique feature of functional and curve data. A natural way to model it is as the action of a group of transformations of the time domain, represented by a class of warp maps. This talk will elaborate on some of the invariant properties of functional data with respect to time warping, and their implications on identifiability of template plus noise models and quality of kernel-based methods.

Monday 22
10:30 - 12:30
Room 5

Detection Thresholds for Non-Parametric Tests Based on Geometric Graphs

Bhaswar Bhattacharya[1]

[1]Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, United States
bhaswar.bhattacharya@gmail.com

Two of the fundamental problems in non-parametric statistical inference are goodness-of-fit and two-sample testing. These two problems have been extensively studied and several multivariate tests have been proposed over the last thirty years, many of which are based on geometric graphs. These include, among several others, the celebrated Friedman-Rafsky two-sample test based on the minimal spanning tree and the K-nearest neighbor graphs, and the Bickel-Breiman spacings tests for goodness-of-fit. These tests are asymptotically distribution-free, universally consistent, and computationally efficient (both in sample size and in dimension), making them particularly attractive for modern statistical applications.

In this talk, we will derive the detection thresholds and limiting local power of these tests, thus providing a way to compare and justify the performance of these tests in various applications. Several interesting properties emerge, such as a curious phase transition in dimension 8, and a remarkable blessing of dimensionality in detecting scale changes.

Analysing Causal Effect of London Cycle Superhighways on Traffic Congestion

Thursday 25
10:30 - 12:30
Room 6

Prajamitra Bhuyan[1] and Emma Mccoy and Daniel Graham

[1]MathematicsImperial College London, London, United Kingdom
prajamitra.bhuyan@gmail.com

Transport operators have a range of intervention options available to improve or enhance their networks. But often such interventions are made in the absence of sound evidence on what outcomes may result. Cycling superhighways was promoted as a sustainable and healthy travel mode which aims to cut traffic congestion. The estimation of the impacts of the cycle superhighways on congestion is complicated due to non-random assignment of such intervention over the transport network. In this paper, we analyse the causal effect of cycle superhighways utilising pre-innervation and post-intervention information on traffic and road characteristics along with socio-economic factors. We propose a modeling framework based on inverse propensity weighting and outcome regression model. The method is also extended to doubly robust set-up. Simulation results show the superiority of the performance of the proposed method over existing competitors. The method is applied to analyse a real dataset on London transport network and the result would help effective decision making to improve network performance.

Parameter Estimation for Stochastic PDEs using High-Frequency Observations

Wednesday 24
10:30 - 12:30
Room 10

Markus Bibinger[1] and Mathias Trabs

[1]Faculty of Mathematics and Computer Science, Philipps University of Marburg, Marburg, Germany
bibinger@mathematik.uni-marburg.de

We discuss parameter estimation for a parabolic linear stochastic partial differential equation from observations of a solution on a discrete grid in time and space. Focusing first on volatility estimation and assuming a high-frequency regime in time, we provide an explicit and easy to implement method of moments estimator based on power variations of time increments. We develop an asymptotic theory for the case with a bounded spatial domain and Dirichlet boundary conditions as well as for the case with an unbounded spatial domain building upon different probabilistic structures. Our estimator is consistent and admits a central limit theorem. Identifiability and estimation of parameters in the differential operator are addressed under different asymptotic observation regimes.

Wednesday 24
10:30 - 12:30
Room 7

Spectral Theory of the Zigzag Sampler

Joris Bierkens[1] and Sjoerd Verduyn Lunel

[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands
joris.bierkens@tudelft.nl

In recent years piecewise deterministic Markov processes (PDMPs) have emerged as a promising alternative to classical MCMC algorithms. The zigzag sampler is a canonical example of such a method.

After a brief introduction of the zigzag process we will approach this process from a functional analytical point of view. In particular we will explore spectral properties of the process. An interesting phenomenon occurs in case of a symmetric target distribution.

Monday 22
10:30 - 12:30
Room 10

Simultaneous Confidence Bands for the Covariance Kernel of Banach Space Valued Functional Data

Melanie Birke[1] and Christoph Reihl and Hajo Holzmann

[1]Mathematical Institute, University of Bayreuth, Bayreuth, Germany
melanie.birke@uni-bayreuth.de

In functional data analysis the mean, convariance kernel and eigenfunctions play an important role to characterize the properties of the data. An important step is to estimated those terms but even more information is given by confidence regions. There already exist some approaches for constructing confidence regions based on dense as well as sparce observational schemes. But most of the methods rely on the weak convergence in Hilbert spaces. This results in confidence regions which are difficult to interpret. Much better interpretation is possible for simultaneous uniform confidence bands. To this end weak convergence results on the Banach space of continuous functions equipped with the supremum norm are necessary. We develope such a result for the local linear estimator of the covariance kernel constructed from observations on a dense grid with additional observation errors in each grid point. This result can be used to construct uniform confidence bands. Besides the theoretical results we present a simulation study. Because of the difficult structur of the asymptotic distribution and because the asymptotics might not be valid for moderate sample sizes we propose a bootstrap method in this part.

Sketched Learning Using Random Moments

Thursday 25
16:30 - 17:30
Aula Magna
Building 13

Gilles Blanchard[1] and Rémi Gribonval and Nicolas Keriven and Yann Traonmilin

[1]Institute of Mathematics, University of Potsdam, Potsdam, Germany
[1]Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France
gilles.blanchard@math.uni-potsdam.de

We introduce and analyze a general framework for resource-efficient large-scale statistical learning by data sketching: a training data collection is compressed in one pass into a low-dimensional sketch (a vector of random empirical generalized moments) that should capture the information relevant to the considered estimation task. The estimation target is the minimizer of the population risk for a given loss function. An approximate minimizer of the empirical risk is computed from the sketch information only using a constrained moment matching principle. Sufficient sketch sizes to control the statistical error of this procedure are investigated. This principle is applied to different setups: PCA, clustering, and Gaussian mixture Modeling. (Joint work with R. Gribonval, N. Keriven and Y. Traonmilin.)

Controlling FDR while Highlighting Selected Discoveries

Friday 26
09:00 - 11:00
Room 11

Marina Bogomolov[1] and Eugene Katsevich and Chiara Sabatti

[1]Department of Industrial Engineering and Management, Technion, Haifa, Israel
marinabo@technion.ac.il

Modern scientific investigations often start by testing a large number of hypotheses by a False Discovery Rate controlling procedure, in order to identify the hypotheses that are promising for follow-up. In many cases, the set of discoveries is somewhat redundant, and it is subject to a second round of selection, where researchers identify the discoveries that better represent distinct findings for reporting and follow-up. For example, in genetic studies, if several genetic variants in a certain locus are identified as associated with the phenotype of interest, typically only the "lead" variant is reported, representing the entire locus. The guarantees of the FDR control for the initial set do not translate to this subset of reported discoveries. We show that if the rule defining how the discoveries will be filtered can be specified in advance, the Benjamini-Hochberg procedure can be modified to result in a focused set of discoveries with FDR control guarantees. The proposed method allows researchers to curate rejections not only by subsetting, but also by prioritizing. We illustrate our methodology on a phenome-wide association study, where the hypotheses are structured as a tree.

Monday 22
10:30 - 12:30
Seminari B

Bayesian Approach to Deconvolution in Well Test Analysis

Themistoklis Botsas[1] and Jonathan Cumming and Ian Jermyn

[1]Mathematical Sciences, Durham University, Durham, United Kingdom
themistoklis.botsas@durham.ac.uk

In the context of petroleum well test analysis, deconvolution is a methodology used to obtain information about the reservoir system, for example the presence of heterogeneities and boundaries. This information is contained in the reservoir response function, which can be estimated by solving an inverse problem in the well pressure and flow rate measurements.

We use a parameterized response function based on the 'multi-region radial composite' model, which is a solution to the diffusion equation under certain symmetry constraints, known in the petroleum engineering community for its ability to resemble most of the response shapes encountered in practice.

We use an errors-in-variables non-linear Bayesian regression model in order to make inferences about the response function. This allows us to include uncertainty for the independent variables (flow rates and initial pressure), which is essential, since the measurements are usually contaminated with large observational error. We combine the likelihood with a set of flexible priors for the response parameters, and we use Markov Chain Monte Carlo algorithms in order to approximate the posterior distribution.

We validate and illustrate the use of the algorithm by applying it to synthetic and field data sets, using a variety of tools to summarize and visualize the posterior distribution, and to carry out model selection. The results are comparable in quality to the state of the art solution, which is based on the total least squares method, but our method has several advantages: we gain access to meaningful system parameters associated with the flow behaviour in the reservoir; we can incorporate prior knowledge that excludes non-physical results; and we can quantify parameter uncertainty in a principled way by exploiting the advantages of the Bayesian approach.

Tuesday 23
14:00 - 16:00
Room 11

On the Structure of Solutions of Convex Regularization: Gradient TV Minimization and Co

Claire Boyer[1] and Antonin Chambolle and Vincent Duval and Yohann De Castro

[1]Sorbonne Université, Paris, France
claire.boyer@upmc.fr

We establish a general principle which states that regularizing an inverse problem with a convex function yields solutions which are convex combinations of a small number of atoms. These atoms are identified with the extreme points and elements of the extreme rays of the regularizer level sets. As a side result, we characterize the minimizers of the total gradient variation, which was still an unresolved problem. This can be viewed to be in the same vein of the representer theorem in machine learning.

A Systematic Checklist for Causal Assessment of Environmental Observational Studies

Friday 26
09:00 - 11:00
Room 7

Aoibheann Brady[1] and Julian Faraway and Ilaria Prosdocimi

[1]Department of Mathematical Sciences, University of Bath, Bath, United Kingdom
A.brady@bath.ac.uk

Natural environmental systems are highly variable, which can lead to considerable societal impacts, including to safety, well-being and wealth. To make accurate predictions and informed decisions under different conditions, it is vital to understand and accurately attribute the mechanisms driving long-term, large-scale environmental change. However, many standard methods for establishing causality are usually either not appropriate or easily applicable to environmental observational studies. We provide a review of past approaches for assessing evidence of causal relationships, including instrumental variable methods, causal diagrams, methods of multiple working hypotheses, and the Bradford Hill criteria. We assess the fitness for purpose of the these methods for environmental studies and select a number of criteria appropriate to these problems. We also propose further steps to be taken within the analysis to address environment-specific issues, incorporating them into a framework which assesses for causal relationships through an audit of strength of evidence for a broad class of problems. We demonstrate this method through a case study of peak river flows in Great Britain.

Lévy Driven Continuous Time Moving Average and Strong Mixing Processes Sampled at a Renewal Sequence

Tuesday 23
14:00 - 16:00
Room 6

Dirk-Philip Brandes[1] and Imma Valentina Curato and Robert Stelzer

[1]Institute of Mathematical Finance, Ulm University, Ulm, Germany
dirk.brandes@uni-ulm.de

We consider two kind of processes. A Lévy driven continuous time moving average process X and a strong mixing or weak depended process Y sampled at random times which follow a renewal structure independent of X. We show that the strict stationarity property of the process X is inherited by the sampled process. Asymptotic normality of the sample mean and the sample autocorrelation is established and sufficient conditions on the kernel and the random times are given. We give an application to parameter estimation of the generalized Ornstein-Uhlenbeck process. For the process Y, we show that the strong mixing or weak dependence property is preserved under renewal sampling.

Tuesday 23
10:30 - 12:30
Room 8

Bootstrapping forIimpulse Response Function Coefficients in Generalized Linear Dynamic Factor Models

Alexander Braumann[1] and Jens-Peter Kreiß

[1]TU Braunschweig, Braunschweig, Germany
a.braumann@tu-braunschweig.de

We consider a bootstrap method for the coefficients of impulse response functions, which are functions of the parameters of a vector autoregressive model for a static factor process. This factor process itself is an unobserved part of a generalized dynamic factor model as proposed in Forni et al. (2000), Forni and Lippi (2001), Forni et al. (2009). In such models the panel of wide-sense stationary time series can be large. The idiosyncratic components are weakly dependent in their cross-sectional and time dimensions. The static factors need to be estimated before estimating the parameters of interest.

The asymptotic distribution of the impulse response function coefficients in such models has been derived in Yamamoto (2018). We propose a bootstrap method which is asymptotically valid for the case of weak dependence of the idiosyncratic variable both in the cross-sectional and the time dimension. Our work can therefore be seen as an extension to the work of Yamamoto (2019). The asymptotic validity is shown using the asymptotic framework of Bai (2003), Goncalves and Perron (2014) and results from Goncalves and Perron (2018).

Tuesday 23
10:30 - 12:30
Room 12

Robust Shape Inference from a Sparse Approximation of the Gaussian Trimmed Loglikelihood.

Claire Brécheteau[1] and Clément Levrard and Bertrand Michel

[1]Laboratoire Jean Leray, Ecole Centrale de Nantes, Nantes, France
claire.brecheteau@ec-nantes.fr

Given a noisy sample of points lying around some shape M, with possibly outliers or clutter noise, we focus on the question of recovering M, or at least geometric and topological information about M. Often, such inference is based on the sublevel sets of distance-like functions such as the function distance to M, the distance-to-measure (DTM) or the k-witnessed distance. In this talk, we firstly widespread the concept of trimmed log-likelihood to probability distributions. This trimmed log-likelihood can be considered as a generalisation of the DTM.

A sparse approximation of the DTM, the m-power distance-to-measure (m-PDTM), has been introduced and studied by Brécheteau and Levrard in 2017. Its sublevel sets are unions of m balls, with m possibly much smaller than the sample size. By miming the construction of the m-PDTM from the DTM, we propose an approximation of the trimmed log-likelihood associated to the family of Gaussian distributions on R^d. This approximation is sparse is the sense that its sublevel sets are unions of m ellipsoids.

We provide a Lloyd-type algorithm to compute the centers and covariance matrices associated to the ellipsoids. We improve our algorithm by allowing an additional noise parameter to wipe out some points, just as the trimmed-means algorithm of Cuesta-Albertos et al.

Our algorithm comes together with a heuristic to select this parameter. Some illustrations on different examples enhance that our algorithm is efficient in wiping out clutter noise, recovering the shape and recovering the homology of M; this requiring a storage of only m points and covariance matrices.

Statistical Challenges when Analysing Emerging Epidemic Outbreaks

Monday 22
14:00 - 16:00
Seminari B

Tom Britton[1] and Gianpaolo Scalia Tomba

[1]Department of Mathematics, Stockholm University, Stockholm, Sweden
tom.britton@math.su.se

New infectious disease outbreaks have great impact on communities over the world, as recently manifested by the Ebola outbreak. An important statistical task is then to predict the future scenario with and with out preventive measures. In the current talk we will investigate such analyses and see how it can be improved. The main catch is that in the exponentially growing phase early on in an outbreak, several biases can occur if not taken account for: events with short delays will be over-represented. We will give some examples from the Ebola outbreak and see how the biases can be removed or at least reduced. (Joint work with Gianpaolo Scalia Tomba).

Asymptotics for Spherical Autoregressions

Monday 22
10:30 - 12:30
Room 10

Alessia Caponera[1] and Domenico Marinucci

[1]Sapienza University of Rome, Rome, Italy
alessia.caponera@uniroma1.it

In recent years, the analysis of spherical random fields has been strongly motivated by a variety of applications in several different areas, such as cosmology, astrophysics, atmospheric sciences and geophysics. Specifically, many natural phenomena involving, for instance, climate change and atmospheric variables, have raised the interest in the analysis of data distributed over the whole sphere and, also, evolving through time. Here, we present a class of spatio-temporal processes, which can be viewed as functional autoregressions taking values in the space of square integrable functions on the sphere. We exploit some natural isotropy requirements to obtain a neat expression for the autoregressive functionals, which are then estimated by a form of frequency-domain least squares. For our estimators we are able to show consistency and limiting distributions. We prove indeed a quantitative version of the central limit theorem, thus deriving bounds on normal approximations in the Wasserstein distance, with explicit rate of convergence. Our results are then illustrated by simulations.

Wednesday 24
10:30 - 12:30
Room 8

Public Support for an EU-Wide Social Benefit Scheme: Evidence from Round 8 of the European Social Survey (ESS)

Paolo Emilio Cardone[1]

[1]INAPP, SAPIENZA University of Rome, Rome, Italy
p.cardone@inapp.org

In the past decades, the extended European-style welfare state became substantially challenged due to a number of major economic, social and political developments. Longer-term challenges have been exacerbated by the shock of the banking crisis in 2008. Many countries have experienced government-imposed austerity measures since then, and many areas of public expenditure have been stagnant, scaled back or cut completely. Finally, there is an ongoing European Union (EU) debate, ignited substantially by the unequal degree to which the economic crisis has hit the different countries in Europe. It regards the solidarity between Europeans, addressing the question of whether a redistribution of welfare from richer to poorer Europeans would be necessary to create cross-European social cohesion, and would be politically and economically feasible. The ESS Round8 module (fielded in 2016/17) - Welfare Attitudes in a Changing Europe: Solidarities under Pressure - makes it possible to shed scientific light on these debates. Using logistic regression model it is possible to estimate the different attitudes among countries for an EU-wide social benefit scheme more accurately in a comparative perspective. The model has been developed for EU citizens only and includes adults' socio-demographic, job and economic characteristics (worked or not, total household income). In order to achieve this goal, we have used "social benefit scheme" as the dependent variable. The Welfare module shows that 67.1

Monday 22
16:30 - 18:30
Room 8

Four Factor Model in Italian Equities Market

Paolo Emilio Cardone[1] and Alessandra Cannarsa

[1]INAPP, SAPIENZA University of Rome, Rome, Italy
p.cardone@inapp.org

We compute the Fama-French and momentum factor returns for the Italian equity market for the January 2008 - January 2019 period using data from Datastream. We cover a great number of firms of the Italian Market and we have classified firms into small and big using a more appropriate cut-off considering the distribution of firm size. This methodology involves a cross classification of stoks on two dimensions: size, measured by market capitalization (Big/Small), and value, measured by the ratio of book value per share to market price per share (High/Medium/Low). For the value breakpoints, we created the High value group, H, consisted of the top 30% stocks in terms of the B/M ratio and the Low stocks (low value group), L, comprised the bottom 30% stocks in terms of the B/M ratio. Following the literature, the Fama-French factors – size and value – were computed using the six disaggregated portfolios (BH, BM, BL, SH, SM and SL) and not directly from the five aggregated portfolios (B, S, H, M and L). The classification of stocks as Winners (W) and Losers (L) was done based on their momentum returns at the end of each month. The momentum returns at the end of month t is the 11-month returns from the end of month t-12 to t-1. The stocks were grouped as W (the top 30% by the momentum return) and L (group consisted of the bottom 30%). Based on the size and momentum groups, four

size-momentum portfolios (WS, WB, LB, LS) were formed every month. The momentum factor WML (Winners minus Losers) was computed as the simple average of the differences in the returns of (WS-LS) and (WB-LB). The WML factor was thus designed to capture the effect of value while being largely free of the influence of size. Finally, we compute the market risk premium as the difference between the return of the market portfolio rate Rm and the Risk-free rate Rf. The time series of daily, monthly and yearly returns on the factors and the underlying portfolios are made available at an

---

Nonparametric Semisupervised Classification and Variable Selection for New Physics Searches

Thursday 25
10:30 - 12:30
Room 7

Alessandro Casa[1] and Giovanna Menardi

[1]Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padova, Italy
casa@stat.unipd.it

Model-independent searches in particle physics look for new and not already predicted particles, referred to as signal, behaving as collective anomalies with respect to the background, representing in turn established physical theories. In the quest for a signal, information from both experimental data and from the background can be used, hence the framework can be recasted to a semisupervised classification problem. In this work we semisupervise nonparametric clustering by considering the information coming from known theories as an aid to estimate the density underlying experimental data. Furthermore we develop a variable selection procedure where the concept of informativeness of features has been specified coherently with the subsequent goal of identifying a possible signal.

Plackett Models for Dependent Discrete Data

Friday 26
09:00 - 11:00
Room 7

Manuela Cattelan[1] and Cristiano Varin

[1]Department of Statistical Sciences, University of Padova, Padova, Italy
manuela.cattelan@unipd.it

Likelihood inference in marginal models for dependent discrete data is difficult because it requires the specification of the multivariate distribution of the observations. We extend the class of bivariate Plackett distributions, typically used for binary data, to other type of discrete data, as for example count data, through the combination with Poisson or negative binomial marginals. This model allows a flexible dependence structure (both positive and negative association) and the dependence parameter is approximately orthogonal to the marginal distributions. Moreover, computation of bivariate distributions is not problematic, hence we suggest to employ pairwise likelihood (Varin, Reid and Firth, 2011) for inferential purposes in order to avoid the specification of the whole multivariate distribution. Simulation studies are performed in order to investigate the finite sample properties of the estimator and the proposed methodology is illustrated through applications to real data sets.

Monday 22
10:30 - 12:30
Room 11

Bootstrapping Non-Stationary Stochastic Volatility

H. Peter Boswijka and Giuseppe Cavaliere[1] and Iliyan Georgiev and Anders Rahbek

[1]Department of Economics, University of Bologna, Bologna, Italy
giuseppe.cavaliere@unibo.it

It is a well established fact that the volatility of many economic and financial time series displays persistent changes and possible non stationarity. However, the theory of the bootstrap for such time series has focused on deterministic changes of the unconditional variance and little is known about the performance and the validity of the bootstrap when the volatility is driven by a non-stationary stochastic process. This includes near-integrated exogenous volatility processes (as analyzed by Hansen, 1995), as well as near-integrated GARCH processes, where the conditional variance has a diffusion limit (Nelson, 1990). This paper aims at filling this gap in the literature by developing conditions for bootstrap validity in time series models with non-stationary, stochastic volatility. We show that in such cases the distribution of bootstrap statistics (conditional on the data) is random in the limit. Consequently, the conventional approach, based on the notion of weak convergence in probability of the bootstrap statistic, fails to deliver the required result. Instead, we use the concept of weak convergence in distribution, see Cavaliere and Georgiev (2017) for a general introduction, to develop novel conditions for validity of the wild bootstrap, conditional on the volatility process. We apply our results to several testing problems in the presence of non-stationary stochastic volatility, including testing for unit roots, inference for predictive regression models and cointegration. Importantly, we show that sufficient conditions for conditional wild bootstrap validity include the absence of statistical leverage effects, i.e., correlation between the error process and its future conditional variance. The results of the paper are illustrated using Monte Carlo simulations, which indicate that the wild bootstrap leads to size control even in small samples.

Tuesday 23
16:30 - 17:30
Second Floor

Hold your Breath! Mathematical Approach to the Influence of Smog on the Respiratory System.

Bożena Cegiełka[1] and Barbara Jasiulis-Gołdyn

[1]Mathematical Institute, University of Wrocław, Wrocław, Poland
bozena.cegielka@wp.pl

Smog is a serious problem in most big urban areas. We rarely realize the consequences of being in polluted air, while mixture of air pollutants can seriously endanger human health. Bronchitis, pneumonia and asthma are only some of the respiratory diseases that are associated with the effects of smog. Polluted air also makes it difficult for people to breathe properly. We analyze the relationship between respiratory diseases and smog based on data from Wroclaw (Poland) regarding calls for ambulance services and indicators of air pollution and meteorological data. We will prepare, among other things, a correlation analysis, based on which we will select significant variables. The next step is to create an econometric model that can be used to predict the number of ambulance calls on a given day. To learn more about this topic, we encourage you to read the poster.

Shannon Entropy Estimation via Discovery Rates: Linking Frequentist and Bayesian Solutions

Monday 22
10:30 - 12:30
Seminari B

Annalisa Cerquetti[1]

[1]Rome, Italy
annalisa.cerquetti@gmail.com

Moving from a result in Chao et al. (2013), we derive infinite series representations of prior and posterior expectation of Shannon entropy in terms of discovery rates under random discrete distributions. The result is introduced through the two-parameter Poisson-Dirichlet prior case and then generalized to the whole Gibbs-type family by means of results on posterior moments of conditional sampling formula (Cerquetti, 2013). It turns out that the best classical solution in terms of bias reduction and the modern Bayesian nonparametric estimator both share the same two-components structure.

Estimation of Breakpoints for Extended Interval Regression Models

Tuesday 23
10:30 - 12:30
Room 5

Chihhao Chang[1] and Kamfai Wong and Weiyee Lim

[1]Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan
jhow@nuk.edu.tw

In data analysis, we are usually interested in analyzing the underlying trend of the data to obtain the important features of the trend. In general, the trend is assumed as a smooth function with a critical point. In this talk, we assume that the critical point can be extended to an interval such that the function has optimal value within this interval with some treatments. We call this model the extended interval regression model and fit the model by a two breakpoints linear regression model having constant values within the two breakpoints. We also establish the asymptotic properties of the breakpoint estimators under some regularity conditions.

Nonlinear Factor Models for Network and Panel Data

Thursday 25
10:30 - 12:30
Room 11

Mingli Chen[1] and Ivan Fernandez-val and Martin Weidner

[1]Economics, University of Warwick, Coventry, United Kingdom
mlchen@bu.edu

Factor structures or interactive effects are convenient devices to incorporate latent variables in panel data models. We consider fixed effect estimation of nonlinear panel single-index models with factor structures in the unobservables, which include logit, probit, ordered probit and Poisson specifications. We establish that fixed effect estimators of model parameters and average partial effects have normal distributions when the two dimensions of

the panel grow large, but might suffer of incidental parameter bias. We show how models with factor structures can also be applied to capture important features of network data such as reciprocity, degree heterogeneity, homophily in latent variables and clustering. We illustrate this applicability with an empirical example to the estimation of a gravity equation of international trade between countries using a Poisson model with multiple factors.

Tuesday 23
16:30 - 17:30
Second Floor

Impact of Multiple Imputation Strategies on the Misspecified Random Effects in Generalized Linear Mixed Models for Incomplete Longitudinal Data

Yi-Ju Chen[1]

[1]Statistics, Tamkang University, New Taipei City, Taiwan
ychen@stat.tku.edu.tw

The analysis of longitudinal data is pervasive in a variety of fields, and generalized linear mixed models are often employed with the normal distribution of the random effects. Sometimes the normality assumption of the random effects is unrealistic for practical application, and the misspecification may result in invalid inference of the fixed effects. To detect the misspecified random effects for binary and ordinal responses, the tests proposed by Huang (2009) and Lin and Chen (2015), respectively, use the difference of the maximized likelihood estimators between the observed data and the reconstructed data. However, the reconstructed data are created by completed data without missing values. Based on simulation studies, we reveal the impact of various multiple imputation strategies under different missingness mechanisms. Furthermore, we demonstrate how to evaluate the adequacy of distributional assumption of the random effects by an example.

Tuesday 23
16:30 - 17:30
Second Floor

A Comparison of Geographically Weighted Regression Models for Count Data

Vivian Yi-Ju Chen[1] and Tzai-hung Wen and Hsin-yu Chu

[1]Department of Statistics, Tamkang University, New Taipei City, Taiwan
viviyjchen@gms.tku.edu.tw

The past years have experienced growth in the methodological development that intend to explore spatial nonsationarity for spatially count data based on the technique of geographically weighted regression. Several geographically weighted count models (GW count modes) have been introduced in literature to deal with the challenges of analyzing the count data with overdispersion and excessive zeros; however researchers have lagged to provide a comparative assessment across all the proposed methods. In this paper, we argue that spatial analysts should pay sufficient attention to analytical model comparisons as different GW count models may generate competing accounts of the same data set. Here we suggest several qualitative measures and graphical tools to compare among various GW count models which include the geographically weighted zero-inflated negative binomial

model introduced as a methodological complement. We also illustrate their utility using an example from a study of Taiwan dengue data. Our results demonstrate the importance of model comparisons in investigating spatial nonstationarity for spatial count data analyses.

Recurrent Neural Networks for ARMA Model Selection

Thursday 25
14:00 - 16:00
Room 7

Bei Chen[1] and Beat Buesser

[1]IBM Research, Dublin, Ireland
beichen2@ie.ibm.com

Selecting an appropriate ARMA model for a given time series is a classic problem in statistics that is encountered in many applications. Typically, this involves a human-in-the-loop and repeated parameter evaluation of candidate models, which is not ideal for learning at scale. We propose a Long Short Term Memory (LSTM) classification model for automatic ARMA model selection. Our numerical experiments show that the proposed method is fast and provides better accuracy than the traditional Box-Jenkins approach based on autocorrelations and model selection criteria. We demonstrate the application of our approach with a case study on volatility prediction of daily stock prices.

High-Dimensional Local Linear Regression under Sparsity and Convex Losses

Monday 22
14:00 - 16:00
Room 8

Kin Yap Cheung[1] and Stephen M.S. Lee

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
u3507181@connect.hku.hk

Existing literature on variable selection for nonparametric regression focus on the mean regression. The application of these methods is restricted by data type of the response. We are going to present a novel penalised local linear regression procedure in a sparse model setting under a wide range of commonly used convex loss function, including the check function and negative log-likelihood function of exponential distribution. The proposed method regularizes bandwidths and regression coefficients in local linear regression using a dual optimization scheme. Both the number of predictors and size of the active set are allowed to grow slowly with sample size which is rarely considered in the existing literature. The data-driven bandwidths selected by the optimization problems help to select variables and achieve an optimal rate of convergence in an oracle sense. Simulation and real data examples of three types of regression, mean regression, quantile regression and logistic regression are presented to show the selection and estimation performance of the new method.

Monday 22
10:30 - 12:30
Room 9

Adaptive Tuning Of Hamiltonian Monte Carlo Within Sequential Monte Carlo

Nicolas Chopin and Alexander Buchholz and Pierre Jacob

[1]ENSAE, CREST, Paris, France
nicolas.chopin@ensae.fr

Sequential Monte Carlo (SMC) samplers form an attractive alternative to MCMC for Bayesian computation. However, their performance depends strongly on the Markov kernels used to rejuvenate particles. We discuss how to calibrate automatically (using the current particles) Hamiltonian Monte Carlo kernels within SMC. To do so, we build upon the adaptive SMC approach of Fearnhead and Taylor (2013), and we also suggest alternative methods. We illustrate the advantages of using HMC kernels within an SMC sampler via an extensive numerical study.

Monday 22
10:30 - 12:30
Seminari B

Nonparametric Posterior Contraction Rates for Discretely Observed Compound Poisson Processes

Alberto J. Coca[1]

[1]Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom
alberto.j.coca@statslab.cam.ac.uk

Compound Poisson processes (CPPs) are a simple, yet fundamental, model for random shocks in a system used in a myriad of problems within natural sciences, engineering and economics. Indeed, they are the textbook example of pure jump Lévy processes and they are dense in this richer class of stochastic processes. The so-called Lévy jump distribution, N, characterises them as it determines the frequency at which jumps (randomly) occur and their (random) sizes.

In most applications, the underlying CPP is not perfectly observed: only discrete observations over a finite-time interval are available. Thus, the process may jump several times between two observations and we are effectively observing a random variable corrupted by a sum of a random number of copies of itself. Consequently, estimating N is a nonlinear statistical inverse problem.

In the recent years, understanding the asymptotic behaviour of the Bayesian method in inverse problems and, in particular, in the problem described above has received considerable attention. In this talk, we will present some recent results on posterior contraction rates for the (multivariate) density ν of N: we show two-sided stability estimates between ν and its image through the forward operator that allow to use the classical theory in Ghosal, Ghosh, van der Vaart (2000) and its subsequent extensions in full generality; furthermore, our results are robust to the observation interval, i.e. optimal adaptive inference can be made without specification of whether the regime is of high- or low-frequency; and, lastly, we propose an efficient ∞-MCMC procedure to draw from the posterior using mixture and Gaussian priors.

Polynomial-Time Estimation of the Mean

Monday 22
16:30 - 18:30
Room 12

Olivier Collier[1] and Arnak Dalalyan

[1]Université Paris-Nanterre, Crosne, France
oliviercollier8@gmail.com

We consider independent Gaussian observations in high dimension, sharing the same mean for the most part, while the other means can be arbitrarily large. We studied the problem of robustly estimating the common mean in the minimax sense. But more precisely, we aim at finding feasible procedures, i.e. computable in polynomial time. First, we showed the relation between robustly estimating the mean and estimating some linear functionals of the outliers. Then, we defined a group-LASSO-like procedure for estimating the mean, which has better performance as previously existing methods. However, computational tractability comes with a loss of minimax-rate-optimality in some regimes.

Testing Quantile Regression Models when the Response Variable is Right Censored

Monday 22
14:00 - 16:00
Room 6

Mercedes Conde Amboage[1] and Ingrid Van Keilegom and Wenceslao Gonzalez-Manteiga

[1]Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Santiago de Compostela, Spain
mercedes.amboage@usc.es

Several authors have addressed the problem of testing a parametric quantile regression model with complete data. However, little can be found in the literature for lack-of-fit testing adapted to censored data in the framework of quantile regression. It should be mentioned the proposal of Wang (2008, Canadian Journal of Statistics) that presents a nonparametric test for checking the lack-of-fit test of a quantile regression model under random right censoring.

Taking into account the state of the art, we propose a new lack-of-fit test for parametric models of quantile regression with censored data versus non parametric alternatives. The test is based on the cumulative sum of residuals, that is, extends the ideas of He and Zhu (2003, Journal of the American Statistical Association) to censored quantile regression. It will be shown that the empirical process associated with the test statistic converges to a Gaussian process under the null hypothesis, and the distribution under alternative hypothesis is also shown. Moreover, to approximate the critical values of the new lack-of-fit test, a bootstrap mechanism is used following the ideas of Li and Datta (2001, Annals of the Institute of Statistical Mathematics).

The performance of the new test is investigated thanks to a Monte Carlo study. This simulation study has shown the good adjustment of the significance level of the test under homo- and heteroscedastic models, with different error distributions and quantiles of interest. In addition, the proposed test is generally more powerful than its natural competitor available in the literature. Finally, a real data application will be presented to show the good properties of the new lack-of-fit test in practice.

Monday 22
10:30 - 12:30
Seminari B

Relative Privacy Threats and Learning from Anonymized Data

Fabio Corradi[1] and Michele Boreale and Cecilia Viscardi

[1]Dipartimento Statistica, Informatica, Applicazioni, Firenze, Italy
fabio.corradi@unifi.it

We consider group-based anonymization schemes, a popular approach to data publishing. This approach aims at protecting privacy of the individuals involved in a dataset, by releasing an *obfuscated* version of the original data, where the exact correspondence between individuals and attribute values is hidden. When publishing data about individuals, one must typically balance the *learner*'s utility against the risk posed by an *attacker*, potentially targeting individuals in the dataset. Accordingly, we propose a unified Bayesian model of group-based schemes and a related MCMC methodology to learn the population parameters from an anonymized table. This allows one to analyze the risk for any individual in the dataset to be linked to a specific sensitive value, when the attacker knows the individual's nonsensitive attributes, beyond what is implied for the general population. We call this *relative threat* analysis. Finally, we illustrate the results obtained with the proposed methodology on a real-world dataset.

Thursday 25
10:30 - 12:30
Room 7

A Novel Approach for Pre-Filtering Event Sources using the Von Mises Distributions

Denise Costantin[1] and Giovanna Menardi and Alessandra Brazzale and Junhui Fan

[1]Guangzhou University, Ghuangzhou, China
denise.costantin@gmail.com

Astronomical observations associate with galactic or celestial coordinate are spherically mapped data, with location projected onto a conceptual sphere. Several specific data structure are developed to store and access spherically mapped data. In general, they use a hierarchical multi-resolution tessellation of the sky. In a clustering context, the analysis of directional data, distributed on the unit sphere surface, rely on the family of directional distributions. We explore the capabilities of a model-based clustering, using a mixtures of Von Mises-Fisher (vMF) distribution, that are natural for directional data, applied in an astrophysical content, with a specific application to Fermi Large Area Telescope data set.

Optimal Designs in Post-Genome-Wide Association Two-Phase Sampling Studies

Monday 22
14:00 - 16:00
Room 7

Radu Craiu[1] and Osvaldo Espin-garcia and Shelley Bull

[1]Statistical Sciences, University of Toronto, Toronto, Canada
craiu@utstat.toronto.edu

High costs in next-generation sequencing limit post-GWAS analysis in large cohorts. Two-phase sampling design (TPD) utilizes data from phase 1 GWAS to inform a subsample in phase 2 (PH2). The goal is to make inference on an expensive, missing-by-design sequence variant (SV) by combining data from phases 1 & 2. Techniques to optimally select the PH2 subsample under budget constraints are scant beyond binary outcomes. Considering as inputs the effect size and haplotype distribution between a GWAS-auxiliary-variable and a SV, we develop two approaches for optimal TPDs under a semiparametric maximum likelihood framework, (1) a Laplace multiplier (LM) method that allows an analytical expression for the variance-covariance matrix (VCM) subject to a budget constraint, and (2) a genetic algorithm (GA) that performs a direct search of the PH2 sampling space. We evaluate optimality criteria based on the VCM useful in experimental design. We perform comprehensive simulations to assess the properties of LM and GA. Our findings suggest that the optimal TPDs can render higher power in combined phase 1 & 2 analysis compared to intuitive designs while preserving type 1 error.

On Smoothing Methods for Circular Regression with Categorical Data

Wednesday 24
10:30 - 12:30
Room 11

Rosa M. Crujeiras[1] and María Alonso-Pena and Jose Ameijeiras-Alonso

[1]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Santiago de Compostela, Spain
rosa.crujeiras@usc.es

Flexible approaches to regression modeling with circular covariates and/or responses have been analyzed by Di Marzio et al. (2009, 2013). When a categorical covariate may influence the response, an ANCOVA model can be constructed. This work will present some new proposals to deal with ANCOVA models in circular regression, that is, regression models where the response and/or the (non categorical) covariate is circular. Following the ideas of Young and Bowman (1995), test statistics for the equality of curves among groups and test statistics for parallelism will be introduced. Their finite sample performance will be analyzed through an exhaustive simulation study, calibrating the distribution under the null hypothesis by bootstrap methods. Some real data application will be also provided for illustration.

Wednesday 24
10:30 - 12:30
Seminari B

A New Regression Model

Valentina Cueva López[1] and José Rodríguez-Avi and María José Olmo-Jiménez

[1]Statistics and Operation Research, University of Jaén, Jaén, Spain
vcueva@ujaen.es

The extended biparametric Waring (EBW) distribution is a biparametric distribution which can be seen, in part, as a version of the generalized univariate Waring (UGW) distribution when the first two parameters are the same. Consequently, the EBW distribution inherits the properties of the UGW distribution, among them, the suitability for modelling overdispersed count data and the explanation of the variance in terms of randomness, liability and proneness. Moreover, the EBW distribution is also useful to model underdispersed count data. In fact, when the first parameter is negative, the EBW distribution may be considered as a particular case of the complex triparametric Pearson (CTP) distribution. Thus, this distribution is an adequate alternative for both under- and overdispersed count data. Taking advantage of this property, in this work we propose a regression model for count data based on the EBW distribution. Then, we develop a generalized lineal regression model where the coefficient estimates are obtained by the maximum likelihood method. We illustrate the performance of the regression model proposed with some real examples and we compare the fits obtained with those obtained by some other usual regression models for count data using the Akaike Information criterion (AIC).

Monday 22
10:30 - 12:30
Room 11

Towards a General Theory for Non-Linear Locally Stationary Processes

Rainer Dahlhaus[1] and Stefan Richter and Wei Biao Wu

[1]Department of Mathematics/Institute of Applied Mathematics, Heidelberg University, Heidelberg, Germany
dahlhaus@statlab.uni-heidelberg.de

Some general theory is presented for locally stationary processes based on the stationary approximation and the stationary derivative. Strong laws of large numbers, central limit theorems as well as deterministic and stochastic bias expansions are proved for processes obeying an expansion in terms of the stationary approximation and derivative. In addition it is shown that this applies to some general nonlinear non-stationary Markov-models. In addition the results are applied to derive the asymptotic properties of maximum likelihood estimates of parameter curves in such models. The approach is also used to derive results on adaptive bandwidth selection via cross validation for local M-estimators in locally stationary processes.

Dimensional Reduction of a Quality of Life Questionnaire: Is Confirmatory Factor Analysis a Powerful Tool?

Tuesday 23
16:30 - 17:30
Second Floor

Nadia Dardenne[1] and Anne-Lise Poirrier and Clotilde De Dorlodot and Anne-françoise Donneau

[1]Department of Public Health, University of Liege, Liège, Belgium
ndardenne@uliege.be

Introduction:Most of quality of life questionnaires (QOLQ) consist of subdomain structures that are explored or clinically identified but not systematically statistically validated. However, validation is required to compare results from various studies. The aim of this study was to evaluate the ability for confirmatory factor analysis (CFA) to validate several models of a specific QOLQ. Their discriminant power was then assessed for distinct groups of patients. Methods:A total of 383 patients with rhinological complaints were recruited in 3 Belgian rhinology clinics. The 22-item Sino-Nasal Outcome Test (SNOT-22) was measured for each patient divided into 3 groups: medical condition, nose or sinus surgery. Exploratory factor analysis was first performed to determine SNOT-22 subdomain structure using polychoric correlation and PROMAX rotation. CFA was then carried out to validate the explored, a previous validated and a commonly accepted model using the weighted least squares means and variance estimator. Their fit was compared using root mean square error of approximation (RMSEA), comparative fit index (CFI), standardized root mean square residual (SRMR) and Tucker-Lewis index (TLI). The discriminant power of each model was then studied using quadratic discriminant analysis with 2-fold cross validation. Results:Despite different number of factors, models had similar subdomain structures and showed adequate fit with CFI values between 0.98 and 0.99, TLI between 0.97 and 0.99, RMSEA between 0.052 and 0.10 and SRMR between 0.058 and 0.086. The percentage of well-classified patients ranged between 52Conclusion:CFA failed to highlight only one model with a good fit and they all presented a low discriminant power. Future investigations should be made to take into account the effect of potential confounding factors.

Noncausal Vector AR Processes with Application to Economic Time Series

Monday 22
10:30 - 12:30
Room 11

Richard A. Davis[1] and Li Song

[1]Department of Statistics, Columbia University, New York, USA
davis.richarda@gmail.com

Inference procedures for noncausal autoregressive (AR) models have been well studied and applied in a variety of applications from environmental to financial. For such processes, the observation at time t may depend on both past and future shocks in the system. Since noncausal models, especially those with heavy-tails, can exhibit locally explosive behavior, they have been suggested for describing and modeling bubbles in financial time series. In this talk, we will consider extensions of the univariate noncausal AR models to the vector AR (VAR) case. The extension presents several interesting challenges since even a first-order VAR can possess both causal and noncausal components. Assuming a non-Gaussian distribution for the noise, we show how to compute an approximation to the likelihood

function. Under suitable conditions, it is shown that the maximum likelihood estimator (MLE) of the vector of AR parameters is asymptotically normal. The estimation procedure is illustrated with a simulation study for a VAR(1) process and with two macro-economic time series. Semiparametric attempts at estimation for these models will also be mentioned. (This is joint work with Li Song and Jing Zhang.)

Wednesday 24
10:30 - 12:30
Room 5

Regularization Methods for Stochastic Differential Equations

Alessandro De Gregorio[1]

[1]Department of Statistical Sciences, "Sapienza" University of Rome, Rome, Italy
alessandro.degregorio@uniroma1.it

In this talk, we introduce some shrinkage estimators for sparse statistical models and study their asymptotic properties. These regularization methods allow to select a parsimonious model by means of suitable L^q penalties. We also discuss another approach based on the elastic-net procedure which takes into account a combination of the L^1 and L^2 constraints. We apply our results to the penalized estimation of stochastic differential equations.

Wednesday 24
10:30 - 12:30
Room 7

The Bouncy Particle Sampler and Randomized HMC

George Deligiannidis[1] and Daniel Paulin and Alex Bouchard-Côté and Arnaud Doucet

[1]Department of Statistics, Oxford, United Kingdom
deligian@stats.ox.ac.uk

I will discuss some recent results on PDMCMC and in particular on the Bouncy Particle Sampler (BPS). In particular I will show that as the dimension of the target grows, at least for targets that factorise or satisfy some other weak dependence assumption, any finite collection of location and momentum coordinates of BPS converge weakly to the corresponding Randomized Hamiltonian Monte Carlo (RHMC). This is essentially a piecewise deterministic version of the well known HMC algorithm and therefore this establishes a close link between BPS and Hamiltonian dynamics. Next, I will go through some methods for obtaining dimension free convergence rates for RHMC and discuss its implications on the computational cost of BPS. This is joint work with D. Paulin, A. Bouchard-Côté and A. Doucet.

Bivariate Parametric Model for Survival Data under Dependent Censoring

Monday 22
14:00 - 16:00
Seminari B

Negera Wakgari Deresa[1] and Ingrid Van Keilegom

[1]ORSTAT, KU Leuven, Leuven, Belgium
negerawakgari.deresa@kuleuven.be

When modeling survival data, it is common to assume that the (log-transformed) survival time (T) is conditionally independent of the (log-transformed) censoring time (C) given a set of covariates. There are numerous situations in which this assumption is not realistic, and a number of correction procedures have been developed for different models. However, in most cases, either some prior knowledge about the association between T and C is required, or some auxiliary information or data is supposed to be available. When this is not the case, the application of many existing methods turns out to be limited. The goal of this paper is to overcome this problem by developing a bivariate parametric model, that is a type of transformed linear model. We show that the association between T and C is identifiable in this model. The performance of the proposed method is investigated both in an asymptotic way and through finite sample simulations. We also develop a formal goodness-of-fit test approach to assess the quality of the fitted model. Finally, the approach is applied to data coming from a study on liver transplants.

A Bayesian Approach to Estimate the Number and Position of Knots for Linear Regression Splines

Friday 26
09:00 - 11:00
Room 6

Gioia Di Credico[1] and Francesco Pauli and Nicola Torelli

[1]University of Trieste, Trieste, Italy
gioia.dicredico@gmail.com

The relationship between a response variable and one or more continuous predictors can be, in many real applications, modeled by using a flexible piecewise representation. Regression spline functions, defined as piecewise polynomials joined in points called knots, offer a viable solution: varying the number and position of knots may lead to extremely different shapes and when estimating the model simple solutions that avoid the risk of overfitting are sought. In order to enhance the fit of the model, an alternative solution could be to consider locations and number of knots as parameters to be estimated, however, this adds several layers of complexity to the estimation problem. In the paper a Bayesian approach is introduced aimed at solving the problem of estimating these two key quantities in semiparametric regression models with linear splines. Spike and slab prior distributions are adopted in order to estimate the presence or absence of knots in possibly overparametrised models. Once the number of knots is selected, the appropriate model can be fitted using an appropriate estimation technique. The method gives also a first guess on knot locations through inspection of their marginal posterior distributions. This turns out to be useful in the initialization step of algorithms given the difficulties in exploring entirely the possibly high dimensional parameter space. The approach is tested on simulated data with different signal to noise ratios and many knots close to each other which are challenging features for this kind of models. The proposed method is also successfully applied in modeling epidemiological data by using a logistic regression with bivariate linear splines. Finally,

the use of spike and slab priors in the context of generalized linear models is discussed: in particular, the impact of the specification of the variances of the two components of the mixture is considered.

Monday 22
10:30 - 12:30
Room 8

Global Test for High-Dimensional Mediation: Testing Groups of Potential Mediators

Vera Djordjilović[1] and Magne Thoresen

[1]Department of Biostatistics, University of Oslo, Oslo, Norway
vera.djordjilovic@medisin.uio.no

Over the last years, we have witnessed an increased interest for causal mediation analysis in genetic epidemiology, genomics, epigenomics, and neuroscience. What these areas of application have in common is that instead of a single variable on a path between an exposure and an outcome, there is a high-dimensional vector. We address the problem of testing whether a high-dimensional vector may act as a mediator between some exposure variable and the outcome of interest. We propose a global test for mediation, which combines a global test with the intersection-union principle. We discuss theoretical properties of our approach and conduct simulation studies which demonstrate that it performs equally well or better than its competitor. We also propose a multiple testing procedure, ScreenMin, that provides asymptotic control of either familywise error rate or false discovery rate when multiple groups of potential mediators are tested simultaneously. We apply our approach to data from a large Norwegian cohort study, where we look at the hypothesis that smoking increases the risk of lung cancer by modifying the level of DNA methylation.

Thursday 25
14:00 - 16:00
Room 5

Inferring Causality by Gaussian Processes

Petar Djuric[1] and Guanchao Feng

[1]Electrical and Computer Engineering, Stony Brook University, Stony Brook, United States
petar.djuric@stonybrook.edu

Understanding causality permeates many areas of science and engineering. In this presentation, we propose a novel and simple method for discovering Granger causality from noisy time series using Gaussian processes. More specifically, we adopt the concept of Granger causality, but instead of using autoregressive models for establishing it, we work with Gaussian processes. We show that information about the Granger causality is encoded in the hyper-parameters of the used Gaussian processes. The proposed approach is first validated on simulated data and then implemented on fetal heart rate and uterine activity signals acquired in the last two hours before delivery. Our results indicate that uterine activity affects fetal heart rate, which agrees with recent clinical studies.

Estimation of the Extrapolation Range Associated with Extreme-Value Models: Application to the Assessment of Sensors Reliability

Thursday 25
14:00 - 16:00
Room 10

Pascal Dkengne Sielenou[1] and Stéphane Girard and Samia Ahiad

[1]MISTI - INRIA, Université Grenoble Alpes, Grenoble, France
pascal.dkengne-sielenou@inria.fr

Extreme-value theory offers a methodology allowing for extrapolation outside the range of available data. However, the statistical reliability of an extrapolated value such as a return level associated with a long return period is strongly related to mathematical assumptions which are difficult to check on data. Therefore, the natural question which arises in practice is: How far is it possible to extrapolate? Recall that the return level $x(T)$ associated with the return period T is the level which is expected to be exceeded on average once every T observations. In this study, we propose a general bootstrap method to estimate the highest return period $T^\star$ based on the generalized extreme-value distribution so that the confidence intervals of return levels associated with all return periods smaller than $T^\star$ are valid and accurate. Considering a sample from a given population, the estimated value of the return level $x^\star = x(T^\star)$ associated with the return period $T^\star$ is called the extrapolation range. The proposed extreme-value based extrapolation procedure is split into two parts depending on whether the available observations can be considered as a sample of block maxima or not. In either case, the proposed extrapolation procedure is validated on a number of simulated samples from some classical distributions and is applied to assess the reliability of sensors embedded in an autonomous car. First, the measurement errors associated with the sensor to be tested is evaluated by comparison with a reference sensor. Then, the proposed procedure is used to estimate the extrapolation range of these errors.

Method of Moment Based Estimation for the Multivariate COGARCH(1,1) Processes

Tuesday 23
14:00 - 16:00
Room 6

Thiago Do Rego Sousa[1] and Robert Stelzer

[1]Department of Mathematics/ Chair of Mathematical Statistics, Technical University of Munich (TUM), Munich, Germany
thiago.sousa@tum.de

For the multivariate COGARCH model, we obtain explicit expressions for the second order structure of the “squared returns” process observed on a discrete-time grid with fixed grid size. Then, we suggest a generalized method of moments estimator for its parameter. Under moment and strong mixing conditions, we show that the resulting estimator is consistent and asymptotically normal. Sufficient conditions for strong mixing, stationarity and identifiability of the model parameter are also discussed. We investigate the finite sample behavior of the estimator in a simulation study.

Thursday 25
14:00 - 16:00
Room 8

Bounding Distributional Errors via Density Ratios

Lutz Duembgen[1] and Richard Samworth and Jon Wellner

[1]Department of Mathematics and Statistics, University of Bern, Bern, Germany
duembgen@stat.unibe.ch

We present some new and explicit error bounds for the approximation of distributions. The approximation error is quantified by the maximal density ratio of the distribution Q to be approximated and its proxy P. This non-symmetric measure is more informative than and implies bounds for the total variation distance. Explicit approximation problems include, among others, hypergeometric by binomial distributions, (generalized) binomial by Poisson distributions, and the Levy-Poincare projection problem.

Thursday 25
14:00 - 16:00
Seminari B

A Machine Learning Approach to Fault Classification in Photovoltaic Systems using Infrared Imagery

Christopher Dunderdale[1] and Warren Brettenny and Chantelle Clohessy and Ernest Van Dyk

[1]Statistics Department, Nelson Mandela University, Port Elizabeth, South Africa
s214046117@mandela.ac.za

As the global energy demand continues to soar, solar energy has become an attractive and environmentally conscious method to meet this demand. This study examines the use of machine learning techniques for image classification in photovoltaic (PV) systems. The use of SIFT descriptors, combined with random forests, are used to classify images of PV modules faults in a South African setting. The successful implementation of this approach has significant potential for cost reduction in fault detection and classification over currently available methods.

Wednesday 24
10:30 - 12:30
Room 7

On PDMP and their Invariant Measure

Alain Durmus[1] and Arnaud Guillin and Pierre Monmarché

[1]ENS Paris-Saclay, Cachan, France
alain.durmus@cmla.ens-cachan.fr

Piecewise Deterministic Markov Processes (PDMP), similarly to diffusion processes, form an important class of Markov processes, which are used to model random dynamical systems in numerous fields. Recently, interest has grown for their use to sample from a target measure. The resulting class of algorithms is referred to as PDMP Monte Carlo (PDMP-MC) methods. To this end, natural questions arise as to the stationarity of the target measure, the ergodicity of the corresponding process and possible bias introduced by the

method. In this talk, we will present general results regarding PDMP processes which are of great interest in the theoretical study of PDMP-MC. First, we introduce a coupling between two PDMPs following the same differential flow which implies quantitative bounds on the total variation between the marginal distributions of the two processes. Second, two results are established regarding the invariant measures of PDMPs. A practical condition to show that a probability measure is invariant for the associated PDMP semi-group is presented. In a second time, a bound on the invariant probability measures in V -norm of two PDMPs following the same differential flow is established. This last result is then applied to study the asymptotic bias of some non-exact PDMP MC methods. All these results will be applied to the Bouncy particle sampler.

Robust Change-Point Detection in Panel Data

Monday 22
16:30 - 18:30
Seminari B

Alexander Dürre[1] and Roland Fried

[1]Statistics, Technische Universität Dortmund, Dortmund, Germany
alexander.duerre@udo.edu

In panel data we observe a usually high number N of individuals over a time period T. If T is large one often assumes stability of the model over time. Different tests have been proposed to check this assumption. They are usually based on a linear or quadratic statistic, which causes unlimited impact of individual observations. A single outlier can therefore spoil the inference. Furthermore, while these methods are optimal under Gaussianity they often behave poorly under heavy tailed data. We therefore present a new robust change-point test which is based on robustly transformed observations. We derive the asymptotic distribution under short range dependence and N, T going to infinity. A small simulation show its usefulness under heavy tailed distributions.

A Wavelet-Based Imputation Approach for Multivariate Time Series

Friday 26
09:00 - 11:00
Room 12

Idris Eckley[1] and Rebecca Wilson and Matt Nunes and Tim Park

[1]Lancaster University, Lancaster, United Kingdom
i.eckley@lancaster.ac.uk

In the age of big data, the demand for novel approaches to modelling time series structure continues to grow. Such series might be collected at high resolution, across numerous variables and may be non-stationary in structure. However, age-old problems such as missing data can still occur within this contemporary setting. It is this problem that we consider in this talk, describing recent work exploring the use of a wavelet-based framework to impute missing data within a non-stationary, multivariate time series setting.

Friday 26
09:00 - 11:00
Room 10

A Simple Yet Rigorous Introduction into Distance Correlation

Dominic Edelmann[1] and Tobias Terzer

[1]Division of Biostatistics, German Cancer Research Center, 69117 Heidelberg, Germany
dominic.edelmann@dkfz-heidelberg.de

In two seminal papers, Székely, Rizzo, and Bakirov introduced the powerful concept of distance correlation as a measure of dependence between random variables. In contrast to Pearson correlation, which only measures linear dependence, distance correlation can detect any kind of dependence including nonlinear or even nonmonotone associations.

During the last years, enormous interest in the theory and applications of distance correlation and distance covariance has arisen. Distance correlation has been applied to inferring nonlinear gene regulatory networks, detecting cancer pathway crosstalk and assessing associations of familial relationships, lifestyle factors and mortatility. Apart from biomedical research, it has been applied to detecting associations in large astrophysical databases and long-range concerted motions in protein.

In this talk, we give a simple, but mathematically rigorous introduction into distance correlation. Important statistical properties are derived using only a minimum of prequisites. Notably, we will derive the consistency and the asymptotic distribution of distance covariance using standard results from U-statistic theory. Apart from that, only standard calculus and some elementary properties of characteristic functions are required.

We conclude the the talk with an outlook on applications of distance correlation in biostatistics.

Thursday 25
10:30 - 12:30
Room 5

Stepwise Model Selection for SDEs in YUIMA

Shoichi Eguchi[1]

[1]Osaka University, Toyonaka, Japan
eguchi@sigmath.es.osaka-u.ac.jp

There are several studies of model selection for stochastic differential equations (SDEs), for example, the contrast-based information criterion for ergodic diffusion processes and the Schwarz type information criterion for locally asymptotically quadratic models. Based on these studies, we implemented the model selection function for SDEs in R package yuima. However, there is room for improvement in this function. In this talk, we will first overview the model selection methods for SDEs and propose the improvements of model selection function.

A Class of Asymptotically Efficient Estimators Based on Sample Spacings

Monday 22
10:30 - 12:30
Room 12

Magnus Ekström[1] and Sherzod M. Mirakhmedov and S. Rao Jammalamadaka

[1]Swedish University of Agricultural Sciences, Umeå, Sweden
Magnus.Ekstrom@slu.se

We consider a general class of estimators based on higher-order sample spacings, called Minimum Power Divergence Estimators (MPDEs). Such estimators are obtained by minimizing approximations of so-called power divergences between distributions in the model and the true underlying distribution, and include the Maximum Product of Spacings Estimator as a special case. The Maximum Likelihood Estimator (MLE) may be derived in a similar way using the Kullback-Leibler divergence. Spacings-based estimators are especially useful when MLEs do not exist. Our results generalize several earlier studies on spacings-based estimation, by utilizing non-overlapping spacings that are of an order which increases with the sample size. The MPDEs are shown to be consistent as well as asymptotically efficient under a fairly general set of regularity conditions. Simulation studies give further support to the performance of these asymptotically efficient estimators in finite samples, and compare well relative to the MLEs. Unlike the MLEs, some of these estimators are also shown to be quite robust under heavy contamination.

Bootstrapping the Empirical Distribution of a Stationary Process with Change-Point

Monday 22
16:30 - 18:30
Seminari B

Farid El Ktaibi[1] and Gail Ivanoff

[1]Mathematics and Statistics, Zayed University, Abu Dhabi, United Arab Emirates
faridelktaibi@zu.ac.ae

When detecting a change-point in the marginal distribution of a stationary time series, bootstrap techniques are required to determine criti- cal values for the tests when the pre-change distribution is unknown. We propose a sequential moving block bootstrap and demonstrate its validity under a converging alternative. Furthermore, we demonstrate that power is still achieved by the bootstrap under a non-converging alternative. These results are applied to a linear process and are shown to be valid under very mild conditions on the existence of any moment of the innovations and a corresponding condition of summability of the coefficients.

Thursday 25
14:00 - 16:00
Room 11

Depth Measures for Partially Observed Functional Data

Antonio Elías[1] and Laura M. Sangalli and Anna Maria Paganoni and R. Jiménez

[1]Statistics Deparment, University Carlos III of Madrid, Madrid, Spain
aelias@est-econ.uc3m.es

The classical literature on Functional Data Analysis (FDA) focus on the analysis of curves that are commonly observed in a continuous and compact domain. However, the presence of data sets where the functions are not completely recorded is becoming more recurrent in real applications and, unfortunately, this issue invalidates many of the methodologies for FDA.

Methodologies for dealing with this issue have been proposed for principal components analysis, clustering, classification and functional reconstruction. However, and up to our knowledge, there is not a suitable depth for this setting, i.e., a measure that allows to order a partially observed functional data set from the center to outwards. Depth measures are an important tool for robust and non-parametric statistics; They are not only useful for ranking data sets, but for visualization, outlier detection, classification, forecasting and missing values imputation, among others.

In this work, we propose a building-block depth definition that allows to incorporate the uncertainty associated with the censoring process. The first block is an integrated functional depth and the second one a weighting function that penalizes the domain where the sample is poorly observed. The validity of our proposal is studied theoretically and tested by simulation under different censoring settings in terms of expected loss, number of missing intervals and statistical relationship with the process of interest. We show that the depth values obtained by our proposal on the partially observed sample are in mean the values obtained from the fully observed data set. Once we have a suitable definition, we are able to unlock well-known techniques that require the use of a depth measure.

As a case study we consider the AneuRisk dataset that has been analysed for evaluating the role of vascular geometry and hemodynamics in the pathogenesis of cerebral aneurysms. Our goal is to discriminate patients with aneurysms in different districts by using the geometric features of the internal carotid artery expressed by its radius profile and centerline curvature. These two variables are partially observed and many analysis had to be restricted to the domain where all the individuals were recorded. We show that, using our definition on the partially observed sample and DD-classifiers, the classification errors obtained are smaller than in other studies.

The problem of ordering a functional data set that is partially observed is addressed. Given a functional integrated depth measure, we propose to incorporate the uncertainty associated to an unobserved censoring process. For doing this, we define a weighting function that penalizes the domain where the sample is poorly observed whereas the new definition does not punish shortly observed data points. The validity of our proposal is tested by simulation below different censoring settings, showing that our modified measures applied on the partially observed sample are in expectation the values obtained from the fully observed data set. We show how different censoring processes in terms of expected loss and number of missing intervals and violation of the Missing-Completely-At-Random assumption affect to the proposal. Once we have a suitable definition, well-known depth based method can be considered for partially observed functional data samples. In particular, we deal with the classification problem of upper and lower patients from Aneurisk65 data set considering the DD-plot.

Simulating from Conditional Distributions

Tuesday 23
10:30 - 12:30
Room 8

Rasmus Erlemann[1] and Gunnar Taraldsen and Bo Henry Lindqvist

[1]Department of Mathematical Sciences, NTNU, Trondheim, Norway
rasmus.erlemann@ntnu.no

We construct a method for drawing simulations from the conditional distribution X, given a statistic value T(X)=t. This can be used in a number of different applications, such as approximating the conditional distribution, goodness of fit hypothesis testing or improving the efficiency of an unbiased estimator by averaging. We also review and compare alternatives, such as the naive sampler and Gibbs sampler. The method is illustrated by examples of uniform, Gamma and inverse Gaussian distributions. We have also provided a real life example, in which the method was applied for deciding if a model is appropriate for the data set.

Wasserstein-2 Bounds in Normal Approximation under Local Dependence

Monday 22
10:30 - 12:30
Room 5

Xiao Fang[1]

[1]Department of Statistics, The Chinese University of Hong Kong, Hong Kong
xfang@sta.cuhk.edu.hk

We obtain a general bound for the Wasserstein-2 distance in normal approximation for sums of locally dependent random variables. The proof is based on an asymptotic expansion for expectations of second-order differentiable functions of the sum. We apply the main result to obtain Wasserstein-2 bounds in normal approximation for sums of m-dependent random variables, U-statistics and subgraph counts in the Erdős-Rényi random graph. We state a conjecture on Wasserstein-p bounds for any positive integer p and provide supporting arguments for the conjecture.

Friday 26
09:00 - 11:00
Room 8

Fuzzy Item Ambiguity Analysis in Psychological Testing and Measurement

Hojjatollah Farahani[1]

[1]Psychology, Tarbiat Modares University, Tehran, Tehran, Iran
Hojjat.Farahani2@gmail.com

Item Ambiguity is a significant factor in psychological testing and measurement. Item Ambiguity is an unavoidable part of psychological testing and assessment .This item statistic has received no attention in classical test theory (CTT) and item response theory (IRT) of psychometrics so far. All items of a psychological test are of the degree of ambiguity. This is able to influence on item discriminant coefficient , test validity, reliability and diagnostic accuracy. It can be a part of the item fairness in the achievement tests as well. Ambiguity of an item is defined as the degree of the perceived fuzziness of the content of that item (Farahani, Wang and Oles, 2018). To determine the item ambiguity of a test, this paper recommended a 5-stage process using fuzzy logic theory. In this paper, this method was presented and illustrated the calculation steps with a numerical example.

Thursday 25
10:30 - 12:30
Room 12

Non-Parametric Probabilistic Regression Modelling of Electricity Demand Data

Matteo Fasiolo[1] and Simon Nicholas Wood and Yannig Goude and Raphaël Nedellec

[1]School of Mathematics, University of Bristol, Bristol, United Kingdom
matteo.fasiolo@bristol.ac.uk

Reliable electricity load forecasts are an essential input for electricity production planning and power grid management. The UK has historically relied heavily on fossil fuel power plants, whose high ramp-up rates made adjusting for forecasting errors easy. As such stations are replaced with less flexible nuclear plants and renewables, the network will become much more reliant on accurate forecasts. New developments are taking also on the demand side, in particular, the deployment of smart meters is increasing the volume, resolution and complexity of the available electricity demand data. This talk will describe how some of these challenges can be tackled using Generalized Additive Models (GAMs), which are non-parametric regression models offering an interesting compromise between flexibility, interpretability and scalability to Big Data. Besides illustrating how recently developed quantile GAM methods (Fasiolo et al, 2018) can be used in the context of short-term load forecasting, we will also cover some current research directions in this area, in particular, GAM methods for forecasting demand extrema (e.g. daily max demand or extreme demand quantiles) and hierarchical/functional regression approaches for handling different levels of aggregation across costumers or data at different spatio-temporal resolutions.

A Dual Process for the Coupled Wright-Fisher Diffusion

Wednesday 24
10:30 - 12:30
Room 8

Martina Favero[1] and Timo Koski and Henrik Hult

[1]Mathematics department, Mathematical statistics division, Royal Institute of Technology (KTH), Stockholm, Sweden
mfavero@kth.se

The coupled Wright-Fisher diffusion is a multi-dimensional Wright-Fisher diffusion for multi-locus and multi-allelic genetic frequencies, expressed as the strong solution to a system of stochastic differential equations that are coupled in the drift, where the pairwise interaction among loci is modeled by an inter-locus selection.

We derive a dual process to the coupled Wright-Fisher diffusion, which contains transition rates corresponding to coalescence and mutation as well as single-locus selection and double-locus selection. The coalescence and mutation rates correspond to the typical transition rates of Kingman's coalescent process. The single-locus selection rate, not only contains the single-locus selection parameters in a form that generalizes the rates for an ancestral selection graph, but it also contains the double-selection parameters to include the effect of the pairwise interaction on the single locus. The double-locus selection rate reflects the particular structure of pairwise interactions of the coupled Wright-Fisher diffusion.

Moreover, in the special case of two loci, two alleles, with selection and parent independent mutation, the stationary density for the coupled Wright-Fisher diffusion and of the transition rates for the dual process are obtained in an explicit form.

The value of the dual process is that it describes the history of a sample, more precisely, the posterior distribution given an allelic configuration at the present time. Knowing the backward dynamics, or a good approximation of them, is what is needed to perform statistical inference given a sample from the present population, e.g. to design suitable proposal distributions in importance sampling algorithms.

Asymptotic Properties of a Random Graph Evolution Model Driven by a Branching Process

Thursday 25
10:30 - 12:30
Room 10

Istvan Fazekas[1] and Bettina Porvazsnyik and Attila Barta and Csaba Noszaly

[1]Faculty of Informatics, University of Debrecen, Debrecen, Hungary
fazekas.istvan@inf.unideb.hu

The well-known preferential attachment network evolution model leads to a scale-free degree distribution. Then the preferential attachment principle was combined by other ideas such as interactions of several nodes. In this paper we study a network evolution model which is based on interaction of nodes and the evolution of the network is driven by a branching process. Using general theorems from the theory of branching processes, we describe the asymptotic behaviour of the network. Our results are extensions of the theorems of Mori and Rokob (2017).

Friday 26
09:00 - 11:00
Room 10

RKHS Testing for Independence under Right-Censored Data

Tamara Fernandez[1] and David Rindt and Dino Sejdinovic and Arthur Gretton

[1]Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom
t.a.fernandez@ucl.ac.uk

Weighted log-rank tests are arguably the most widely used tests by practitioners for hypothesis testing in the context of right-censored data. Many approaches have been considered to make weighted log-rank tests more robust against a broader family of alternatives, among them: considering linear combinations of weighted log-rank tests or taking the maximum among a finite collection of them. In this work, we propose to non-parametrically test independence between survival times and covariates by considering the supremum of a collection of (potentially infinite) weight-indexed log-rank tests where the index space is the unit ball of a reproducing kernel Hilbert space (RKHS). Our test can naturally handle multiple continuous covariates and by using the good properties of the RKHS's we provide an exact and simple evaluation of the test-statistic and establish relationships between previous tests in the literature.

Monday 22
10:30 - 12:30
Room 12

Nonparametric Likelihood Density Estimation on Bounded Domains

Federico Ferraccioli[1] and Laura M. Sangalli and Eleonora Arnone and Livio Finos

[1]Statistical Sciences, Università degli Studi di Padova, Castelnovo Bariano, Italy
federicoferraccioli@gmail.com

A nonparametric penalized likelihood approach for density estimation of point data scattered over complex planar domains is considered. The model is suitable for any type of bounded planar domain, possibly characterized by strong concavities or holes, and can be generalized to non-Euclidean settings. A regularizing term based on partial differential operators, such as the Laplace operator, controls the amount of smoothness of the estimate while preventing degenerate solutions. The infinite-dimensional poblem is discretized by mean of a finite element formulation. This formulation allows great flexibility and computational tractability. Asymptotic results based on M-estimation theory and metric entropy are discussed. In particular, the well-posedness of the considered estimation problem and the consistency of the associated estimator are studied. The performances of the proposed estimator are investigated through simulations studies. Within this framework, a permutation-based procedure for one and two samples hypothesis testing is also introduced.

Ranking the Importance of Genetic Factors by Variable-Selection Confidence Sets

Thursday 25
10:30 - 12:30
Room 12

Davide Ferrari [1]

[1]Faculty of Economics and Management, University of Bolzano, Bolzano, Italy
davferrari@unibz.it

The widespread use of generalized linear models in case–control studies has helped to identify many disease-associated risk factors typically defined as DNA variants, or single-nucleotide polymorphisms (SNPs). Up to now, most literature has focused on selecting a unique best subset of SNPs based on some statistical perspective. When the noise is large compared with the signal, however, multiple biological paths are often found to be supported by a given data set. We address the ambiguity related to SNP selection by constructing a list of models—called a variable-selection confidence set (VSCS)—which contains the collection of all well-supported SNP combinations at a user-specified confidence level. The VSCS extends the familiar notion of confidence intervals in the variable-selection setting and provides the practitioner with new tools aiding the variable-selection activity beyond trusting a single model. On the basis of the VSCS, we consider natural graphical and numerical statistics measuring the inclusion importance of an SNP based on its frequency in the most parsimonious VSCS models.

The dcortools R Package: Fast Algorithms for the Calculation of the Distance Correlation and Applications

Friday 26
09:00 - 11:00
Room 10

Jochen Fiedler[1] and Dominic Edelmann

[1]Mathematics for Vehicle Engineering, Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany
jochen.fiedler@itwm.fraunhofer.de

Simple implementations of the usual empirical distance covariance are of order n^2 and thus too slow and memory intensive to deal with larger vectors. To solve this problem, Huo and Szekely developed estimators based on U-statistics, which are of order $nlog(n)$. The R-package dcortools now provides some functions for the fast calculation of distance covariances using the algorithms developed by Huo and Szekely. Even for large vectors these implementations allow the fast calculation of distance correlations as well as the calculation of independence test statistics. In this talk we present some functions, illustrate their usage and compare their speed with other implementations.

Tuesday 23
10:30 - 12:30
Room 11

Quasi-Likelihood for Compositional Data

David Firth [1] and Fiona Sammut

[1]Statistics, University of Warwick, Coventry, United Kingdom
d.firth@warwick.ac.uk

A general framework is developed for regression modelling of a multivariate compositional response, i.e., a response vector whose elements are non-negative with fixed sum. Central to the development is a quasi-likelihood formulation that extends a variance function due to Wedderburn (1974) to multivariate responses. This quasi-likelihood is shown to be a 'canonical' one for modelling compositional data on their original scale, without transformation of the response variable. The resulting generalized linear models are readily interpretable, and they avoid common difficulties associated with zero-valued or near-zero responses. Applications are in diverse fields, ranging for example from geology (composition of rocks or sediments) to social science (e.g., time-use data).

Tuesday 23
14:00 - 16:00
Room 8

Consistent Scoring Functions and Murphy Diagrams for Set-Valued Measures of Systemic Risk

Tobias Fissler[1] and Jana Hlavinova and Birgit Rudloff

[1]Mathematics, Imperial College London, London, United Kingdom
t.fissler@imperial.ac.uk

We establish elicitability and identifiability results for measures of systemic risk introduced in Feinstein, Rudloff and Weber (SIAM J. Financial Math. 2017). These risk measures determine the set of all capital allocations that make a financial system acceptable. Hence, they take an ex ante angle, specifying those capital allocations that prevent the system from default. At the same time, they allow to capture the dependence structure of different financial firms.

The elicitability of a risk measure, or more generally, a statistical functional, amounts to the existence of a strictly consistent scoring or loss function. That is a function in two arguments, a forecast and an observation, such that the expected score is minimised by the correctly specified functional value, thereby encouraging truthful forecasts. Prominent examples are the squared loss for the mean and the absolute loss for the median. Hence, the elicitability of a functional is crucial for meaningful forecast comparison and forecast ranking, but also opens the way to M-estimation and regression.

To allow for a rigorous treatment of elicitability of set-valued functionals, we introduce two modes of elicitability: a selective and an exhaustive version. We show that these two modes are mutually exclusive and establish exhaustive elicitability results for the systemic risk measures under consideration. That means we construct strictly consistent scoring functions taking sets as input arguments for forecasts.

Our construction relies on a mixture representation of elementary scores akin to the one established for quantiles and expectiles in Ehm, Gneiting, Jordan and Krüger (JRSS B,

2016). This naturally leads to multivariate Murphy diagrams which can become an important diagnostic tool in evaluating forecasts simultaneously with respect to all consistent scoring functions.

A Modification of the Simplicial Depth Function

Tuesday 23
14:00 - 16:00
Room 10

Giacomo Francisci[1] and Claudio Agostinelli and Alicia Nieto-Reyes

[1]Matematica, Trento, Italy
giacomo.francisci@unitn.it

Spherical, elliptical, central, angular and halfspace symmetry are maybe the most popular notions of symmetry in multivariate spaces. We study the conditions under which linear combinations of symmetric random variables under each of these notions are again symmetric. As an incidental product, we propose a modification of the simplicial depth that does not vanish outside the support of the distribution. This modification is based on the enlargement of the simplices used in the definition. We will present some theoretical results as well as simulations that show the nice behaviour of this modified depth.

Stirring the Mix: Negative Association and Permutation Polytopes

Thursday 25
14:00 - 16:00
Room 5

Lorenzo Frattarolo[1] and Roberto Casarin and Radu Craiu and Christian P. Robert

[1]Unit B1, Joint Research Center, Ispra, Italy
lorenzo.frattarolo@ec.europa.eu

Concepts of negative dependence are useful in probability theory, e.g. for deriving inequalities, in statistical modeling, e.g. to represent dependence relations that have certain known properties, or in statistical computation, e.g. to reduce the variance of Monte Carlo estimators. The notion of negative association is important as its impact permeates all three directions. However, proving that a random sequence has the negative association property is difficult outside of a few well-known constructions. We sheds new light and power on older methods devised for building sequences of negative dependent random variables. We focus on computational statistics applications where sequences with negative association are used for reducing the Monte Carlo error.

Wednesday 24
10:30 - 12:30
Room 10

How Simultaneous Confidence Regions Act in Nonlinear Diffusion Models

Claudia Furlan[1] and Cinzia Mortarino

[1]Dept. of Statistical Sciences, University of Padua, Padova, Italy
furlan@stat.unipd.it

Accuracy measures for parameter estimates represent a tricky issue in nonlinear models. Practitioners often use separate marginal confidence intervals for each parameter in place of a simultaneous confidence region (sCR). This can be extremely misleading due to the curvature of the parameter space of the nonlinear model. For low parameter dimensions, routines for evaluating approximate sCRs are available in the most common software programs. Exact sCRs are computationally intensive, and for this reason, in the past, they did not receive much attention. In this work, we perform a comparison among exact, asymptotic exact, approximate sCRs, and marginal confidence intervals. More modern regions based on bootstrap are also examined. Their degree of accuracy is compared with both real data and simulation results. Among nonlinear models, we focus on 2 of the most widespread S-shaped diffusion models for the lifecycle of products and technological innovation, that is, the Bass (BM) and Generalized Bass (GBM) models, with a parameter space dimension varying from 2 to 6. The results show that, as the parameter dimension increases, overlapping among the alternative sCRs reduces especially with high curvature. The approximate sCR shows inadequate values of overlapping with the exact sCR, even for moderate parameter dimension. Bootstrap regions also exhibit good performance in describing the shape of the exact region when curvature is present, but they fail to spread up to its boundary. The coverage probability of each region is assessed with simulations. We observe that the coverage probability of the approximate sCR decreases rapidly, even for moderate parameter dimension, and it is smaller than the nominal level for bootstrap regions. The exact and the asymptotic exact sCRs are substantially equally performant in terms of coverage probability but in real applications, with a short lifecycle stage and high parameter dimension, they showed a low degree of overlap.

Thursday 25
10:30 - 12:30
Room 12

Target Tracking for Contextual Bandits: Application to Demand Side Management

Pierre Gaillard[1], Margaux Brégère, Yannig Goude and Gilles Stoltz

[1]DI ENS/INRIA, Paris, France,
pierre.gaillard@inria.fr

We propose a contextual-bandit approach for demand side management by offering price incentives. More precisely, a target mean consumption is set at each round and the mean consumption is modeled as a complex function of the distribution of prices sent and of some contextual variables such as the temperature, weather, and so on. The performance of our strategies is measured in quadratic losses through a regret criterion. We offer $T^{2/3}$ upper bounds on this regret (up to poly-logarithmic terms)—and even faster rates under stronger assumptions—for strategies inspired by standard strategies for contextual bandits (like LinUCB). Simulations on a real data set gathered by UK Power Networks, in which

price incentives were offered, show that our strategies are effective and may indeed manage demand response by suitably picking the price levels.

Generalized Bayesian Ensemble Modeling: Methodological and Computational Aspects

Tuesday 23
10:30 - 12:30
Room 8

Marta Galvani[1] and Chiara Bardelli and Silvia Figini and Pietro Muliere

[1]Department of Mathematics, University of Pavia, Pavia, Italy
marta.galvani@unipv.it

In this work a novel approach of ensemble decision trees is proposed, introducing a generalization of the nonparametric Bayesian bootstrap. Ensemble techniques are powerful approaches employed to improve predictive performance combining multiple learning models, such as decision trees. Many ensemble models are based on the bootstrap procedure introduced by Efron. As alternative, some works in literature try to apply the Bayesian version of this procedure, introduced by Rubin. These two techniques are proved to be asymptotically and first order equivalent and despite its Bayesian nature no prior knowledge is introduced when the Rubin bootstrap procedure is used. On the other hand, the generalization of the Bayesian bootstrap, introduced by Secchi and Muliere (1996), allows to consider prior opinions thus including observations not already observed in the sample, assuming the posterior distribution of the sample as a Dirichlet process with parameter depending on the prior and on the empirical distribution obtained by the data at hand. In this work a new algorithm, called Generalized Bayesian Ensemble Tree, is designed to include this generalized procedure in ensemble decision trees. Empirical evidence on simulated and real datasets shows that better predictive performance and more stable results with respect to other ensemble techniques based on different bootstrap procedures, can be obtained through GBET algorithm.

On the Expected Extinction Time for the Adjoint Birth-Death Circuit Chans in Random Environments

Wednesday 24
10:30 - 12:30
Room 8

Chrysoula Ganatsiou[1]

[1]University of Thessaly, Volos, Greece
ganatsio@otenet.gr

It is known that the classical birth-death chain is a special case of homogeneous, aperiodic, irreducible Markov chain (discrete-time or continuous-time), where state changes can only happen between neighbouring states. The model's name comes from a common application, the use of such models to represent the current size of a population where the transitions are literal births and deaths (Wilkinson [4]).

By using the circuit-weight representation theory of Markov chains (S. Kalpazidou [1], J. MacQueen [3]) the present work arises as an attempt to investigate the expected extinction time for the corresponding "adjoint" Markov chains describing uniquely the discrete-time birth-death chains by directed circuits and weights in random environments. (Ch. Ganatsiou [2]).

To this direction we consider a discrete-time birth-death chain $(X_n), n \in \mathbb{N}$ (and its "adjoint" chain $(X'_n), n \in \mathbb{N}$, whose transition probabilities $(p_k), k \in \mathbb{N}$ constitute a stationary and ergodic sequence. A realization of this sequence is called a random environment for this chain. Denoting by t_m the expected extinction time before the population becomes equal to zero taking into account that the initial population is equal to m we are trying to find suitable formulae for the tm associated with the corresponding "adjoint" Markov chains $(X_n), n \in \mathbb{N}$, $(X'_n), n \in \mathbb{N}$ through their unique representations by directed circuits and weights in random environments.

References

[1] S. Kalpazidou, Cycle representations of Markov Processes, Springer, New York, NY, USA, 1995.

[2] Ch. Ganatsiou, On cycle representations of discrete-time birth-death processes, 29th European Meeting of Statisticians, Budapest, Hungary, 20-25/7/2013.

[3] J. MacQueen, Circuit processes, Ann. Probab., 9(1981), 604-610.

[4] D. J. Wilkinson, Stochastic Modelling for Systems Biology, Chapman and Hall, 2011.

Tuesday 23
10:30 - 12:30
Room 10

Scalable Multivariate Bayesian Model for Multi-Level Data

Habib Ganjgahi[1] and Chris Holmes

[1]Statistics Department, Oxford, United Kingdom
h.ganjgahi@gmail.com

The International Mouse Phenotyping Consortium (IMPC) is aimed at measuring the phenotypic consequences of knocking out (KO) each gene in the mouse genome in turn. Several thousand measurements are taken on each animal, in procedures ranging from clinical blood chemistry to behavioural phenotypes. As of March 2019, over 60.5 million data points have been collected from 5600 KO lines, and over 7000 phenotypes. The current practice in analysing the IMPC is to use mass univariate approach, evaluating gene knockout effect on each line and phenotype separately which ignores the shared information between different parameters. In this work, we developed a Bayesian multivariate method to analyse IMPC data to enhance power and reduce false positive rate. However, missing data, nuisance covariates like sex, experimenter and the correlation structure in the IMPC, induced by litter sharing and day of experiment, across the n mice, present unique challenges for multivariate modelling. We have developed 2-steps multivariate model where the effect of data structure and overall knockout effect are evaluated in the 1-step using computationally efficient Bayesian linear mixed effect model. The effect of nuisance variables including , litter, day of experiment and covariates are removed at this stage. In the 2-step, residuals are used to find knock out lines that derived the overall genetic effect. Structured sparsity prior are used at this stage where it's assumed that there are lines that affect many procedures and lines that are specific to a procedure.

Inference of Dependence Graphs by Multiple Testing, with Application to Brain Connectivity

Thursday 25
10:30 - 12:30
Seminari B

Irène Gannaz[1] and Sophie Achard and Pierre Borgnat and Marine Roux

[1]ICJ, INSA Lyon, Lyon, France
irene.gannaz@insa-lyon.fr

In neuroscience, the modeling of the brain as a network defined by the interactions between brain regions has shed lights for understanding brain mechanisms, e.g. evolution of diseases or recovering after a coma. The most common approach consists in thresholding the correlation measured between time series. To control this thresholding, we propose to use a multiple testing strategy. We build correlation tests and choose to control the Family Wise Error Rate (FWER) of the multiple comparisons. After showing theoretically that the FWER is asymptotically controlled for any graph structures, we implement our approach using different procedures Bonferroni, Sidack, Boostratp, MaxT. The control is displayed on simulations, using different graph structures inspired by real data examples. Finally, we illustrate the procedure on a real fMRI dataset.

Proper Likelihood Ratio Based ROC Curves

Friday 26
09:00 - 11:00
Room 7

Mauro Gasparini[1] and Lidia Sacchetto

[1]Dipartimento di Scienze Matematiche, Politecnico di Torino, Torino, Italy
mauro.gasparini@polito.it

Everybody writes that ROC curves, a very common tool in binary classification problems, should be optimal, and in particular concave, non-decreasing and above the 45-degree line. Everybody uses ROC curves, theoretical and especially empirical, which are not so. This work is an attempt to correct this incoherent behavior. Optimality stems from the Neyman-Pearson lemma, which prescribes using likelihood-ratio based ROC curves. Starting from there, we give the most general definition of a likelihood-ratio based classification procedure, which encompasses finite, continuous and even more complex data types. We point out a strict relationship with a general notion of concentration of two probability measures. We give some nontrivial examples of situations with non-monotone and non-continuous likelihood ratios. Finally, we propose the ROC curve of a likelihood ratio based Gaussian kernel flexible Bayes classifier as a proper default alternative to the usual empirical ROC curve.

Friday 26
09:00 - 11:00
Room 5

Network Inference for Truncated Gaussian Data.

Anne Gegout Petit[1] and Clémence Karmann and Aurélie Geudin

[1]Inria BIGS, de Lorraine, 54600 Villers-les-Nancy, France
anne.gegout-petit@univ-lorraine.fr

We are interested in inferring the structure of conditional dependencies in the case of Gaussian data zero-inflated by double truncation. We have a Gaussian p-vector X observed through the truncated vector $Y := X * 1_{a \leq X \leq b}$. The objective is to infer the precision matrix of X from observations of Y. To do this, we propose an estimation procedure, that consists of first estimating each of the terms of the covariance matrix by maximum likelihood estimation, and then the precision matrix using the graphical Lasso. We give a theoretical result about the convergence of the estimated precision matrix.

Thursday 25
10:30 - 12:30
Room 7

Dark Matter Interpretations of Gamma-Ray Signals in the Presence of Unknown Astrophysical Backgrounds

Alex Geringer-Sameth[1] and Roberto Trotta

[1]Dept. of Physics, Imperial College London, London, United Kingdom
a.geringer-sameth@imperial.ac.uk

Deep, all-sky gamma-ray observations combined with recent discoveries of nearby dwarf galaxies have set the stage for the potential detection of dark matter annihilation. However, our limited understanding of the gamma-ray background, especially regarding populations of faint astrophysical sources, severely lowers the power of conventional searches, which are usually optimized for a particular background model. I will present a new method to statistically characterize the background near a dark matter target and then use this derived probability distribution to define a significance test. The method maintains sensitivity to the dark matter signal while taking into account the unknown astrophysical sources of contamination.

Friday 26
09:00 - 11:00
Room 10

The Distance Precision Matrix: Computing Networks from Non-Linear Relationships

Mahsa Ghanbari[1] and Julia Lasserre and Martin Vingron

[1]Computational Regulatory Genomics, Max Delbruck Center for Molecular Medicine, Berlin, Germany
mahsa.ghanbari@mdc-berlin.de

Full-order partial correlation, a fundamental approach for network reconstruction, e.g. in the context of gene regulation, relies on the precision matrix (the inverse of the covariance matrix) as an indicator of which variables are directly associated. The precision matrix

assumes Gaussian linear data and its entries are zero for pairs of variables that are independent given all other variables. However, there is still very little theory on network reconstruction under the assumption of non-linear interactions among variables. We propose Distance Precision Matrix, a network reconstruction method aimed at both linear and non-linear data. Like partial distance correlation, it builds on distance covariance, a measure of possibly non-linear association, and on the idea of full-order partial correlation, which allows to discard indirect associations. We provide evidence that the Distance Precision Matrix method can successfully compute networks from linear and non-linear data, and consistently so across different datasets, even if sample size is low. The method is fast enough to compute networks on hundreds of nodes.

Test for Covariate Effects in Conditional Copula Models

Thursday 25
10:30 - 12:30
Room 9

Irène Gijbels[1] and Marek Omelka and Noël Veraverbeke

[1]Department of Mathematics, Leuven, Belgium
irene.gijbels@kuleuven.be

Conditional copulas describe the conditional dependence and the influence that covariates have on the dependence structure between two (or more) variables. Of interest is to test the null hypothesis that the covariates have a specific effect (linear or quadratic effects, partial effects, ...). A special case is the testing problem that the covariates do not effect the dependence structure. In this talk we discuss some recently developed tests, among which an omnibus test that is designed for having power against many alternatives. In the semiparametric framework the marginal distribution functions are estimated using nonparametric kernel techniques and the parametric dependence model is estimated using maximum likelihood estimation. The asymptotic distribution of the test statistic under the null hypothesis is established, and the finite-sample performance of the test is evaluated via a simulation study. Illustrations will real data examples are given.

Bernstein-Von Mises Theorems and Uncertainty Quantification for Linear Inverse Problems

Monday 22
10:30 - 12:30
Seminari B

Matteo Giordano[1] and Hanne Kekkonen

[1]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge
mg846@cam.ac.uk

We consider the statistical inverse problem of approximating an unknown function f from a linear measurement corrupted by additive Gaussian white noise. We employ a nonparametric Bayesian approach with standard Gaussian priors, for which the posterior-based reconstruction of f corresponds to a Tikhonov regulariser with a Cameron-Martin space norm penalty. We prove a semiparametric Bernstein–von Mises theorem for a large collection of linear functionals of f, implying that semiparametric posterior estimation and uncertainty quantification are valid and optimal from a frequentist point of view. The

result is illustrated and further developed for some examples both in mildly and severely ill-posed cases. For the problem of recovering the source function in elliptic partial differential equations, we also obtain a nonparametric version of the theorem that entails the convergence of the posterior distribution to a fixed infinite-dimensional Gaussian probability measure with minimal covariance in suitable function spaces. As a consequence, we show that the distribution of the Tikhonov regulariser is asymptotically normal and attains the information lower bound, and that credible sets centred at the Tikhonov regulariser have correct frequentist coverage and optimal diameter.

Thursday 25
14:00 - 16:00
Room 7

Robust Tests for White Noise and Cross-Correlation

Liudas Giraitis[1] and Violetta Dalla and Peter C.B. Phillips

[1]School of Economics and Finance, Queen Mary University of London, London, United Kingdom
L.Giraitis@qmul.ac.uk

Commonly used tests to assess evidence for the absence of autocorrelation in a univariate time series or serial cross-correlation between time series rely on procedures whose validity holds for i.i.d. data. When the series are not i.i.d., the size of correlogram and cumulative Ljung-Box tests can be significantly distorted. This paper adapts standard correlogram and portmanteau tests to accommodate hidden dependence and non-stationarities involving heteroskedasticity, thereby uncoupling these tests from limiting assumptions that reduce their applicability in empirical work. To enhance the Ljung-Box test for non-i.i.d. data a new cumulative test is introduced. Asymptotic size of these tests is unaffected by hidden dependence and heteroskedasticity in the series. Related extensions are provided for testing cross-correlation at various lags in bivariate time series. Tests for the i.i.d. property of a time series are also developed. An extensive Monte Carlo study confirms good performance in both size and power for the new tests. Applications to real data reveal that standard tests frequently produce spurious evidence of serial correlation.

Thursday 25
14:00 - 16:00
Seminari B

ETrees: A Generalization of Conditional Trees to Mixed-Type Data

Riccardo Giubilei[1] and Tullia Padellini and Pierpaolo Brutti

[1]Sapienza University of Rome, Rome, Italy
riccardo.giubilei@uniroma1.it

Conditional Trees (CTrees) are a special case of recursive binary partitioning models where statistical tests are used in order to achieve unbiased variable selection and to solve the overfitting problem, and that, so far, have been applied only to either nominal or numeric variables. We propose an extension to the case of mixed-type data, where covariates may include functional data, graphs, and persistence diagrams. The testing procedures that in CTrees characterize both variabile selection and stopping criterion are here performed by means of energy statistics. Energy statistics allow to compare variables that need not

to be defined on the same space, thus permitting to simultaneously model mixed-type covariates. This means that the resulting Energy Trees (ETrees) are a general model which can be applied to a number of cases where other models are not viable, with the additional advantage of being strongly based on statistical testing procedures. The results obtained in both simulated scenarios and real-case analyses are promising, and definitely foster further explorations in the area.

Extreme Value Analysis for PM2.5 Air Pollution in Poland

Tuesday 23
16:30 - 17:30
Second Floor

Joanna Gołdyn[1] and Barbara Jasiulis-Gołdyn

[1]Mathematical Institute, University of Wrocław, Wrocław, Poland
joanna.goldyn@op.pl

In this study we focus on the daily maximum level of PM2.5 in Poland. This tiny air pollution (less than 2.5 micrometers diameter) seems to have a crucial influence on human health. Statistical analysis of actual data helps us to understand the phenomenon the data are referring to. By analysing last few years data, its descriptive statistics and trends, we fit models (e.g. GAM, GEV), compare the quality of fitting and select the best one. The next values of considered pollution are predicted using the model. Additionally we check which other air pollution shows the highest correlation with PM2.5.

---

changedetection: An R Package for Nonparametric Structural Change Detection in Multivariate Systems

Friday 26
09:00 - 11:00
Room 8

Olga Gorskikh[1] and Olga Gorskikh and Pekka Malo and Lauri Viitasaari

[1]Department of Information and Service Management, Aalto University, Espoo, Finland
olga.gorskikh@aalto.fi

This paper introduces the R package “changedetection”, which implements the nonparametric structural change point detection method (NSA) discussed in Malo et al. (2018). The method combines the nonparametric energy distance minimization principle introduced by Rizzo and Szekely (2010) with regularized regression techniques, which makes it applicable to large dimensional multivariate multiple regression problems. The performance of the proposed method has been found to be robust under a variety of settings ranging from the usual gaussian case to problems with fat-tailed error distributions, substantial amount of outliers and large number of nuisance variables.

Wednesday 24
10:30 - 12:30
Room 8

A Fault Prediction and Classification Method for Temporal Data

Emanuele Gramuglia[1]

[1]Mathematics Department, University of Oslo, Oslo, Norway
emanueg@math.uio.no

Fault detection techniques have been studied by the statistics community since the late 1970s, encouraged by the fast development of software technologies for monitoring systems. The increased availability of complex and detailed information leads to several kind of data structures involving time, such as data streams, temporal networks and time series data. Between these, event logs are becoming extensively used for monitoring because of their high reliability in determining the health status of the system. Our work provides a new log-based statistical methodology for fault classification and prediction. The model assumes an unobservable process of breakpoints defining patterns within the log file; each pattern is in turn linked to one of several possible states by mean of two key features, the rate of occurence of the events and teir codification. Specifically, a Bayesian Herarichical non-Homogenous Poisson Process is assumed for modelling the occorurence over time of breakpoints and events while the Lantent Dirichlet Allocation model is employed for modelling the composition of the states in terms of the events codification. The flexibility of the model allows simultaneously to cluster the events according to the different phases the system experiences, to learn the composition of each state and classfiy the clusters according their events . Importantly, the prediction of the state of a new cluster is naturally handled by the model by looking at the codification of its events.

Thursday 25
14:00 - 16:00
Room 7

On Partial-Sum Processes of ARMAX Residuals

Steffen Grønneberg[1] and Benjamin Holcblat

[1]Department of economics, BI Norwegian Business School, Oslo, Norway
steffeng@gmail.com

We study partial-sum processes of ARMAX residuals under weak conditions, and identify transformations that leads to asymptotic pivot processes. Examples of the ARMAX models considered are ARMA with potentially a unit root and a polynomial time trend, and an ARMA with a structural break. An application of these results is CUSUM tests for structural stability of the parameters in the model.

Moments of Local Dirac Mixtures - Algebra and Applications in Statistics

Tuesday 23
14:00 - 16:00
Room 5

Alexandros Grosdos Koutsoumpelias[1] and Markus Wageringel

[1]Mathematics Institute, Osnabrück University, Osnabrück, Germany
algrosdoskou@uos.de

Moments and cumulants of statistical distributions have recently gained attention from an algebraic and geometric point of view. In this talk we introduce local mixtures of Dirac measures and their moment ideals. Adding to the statistics-to-geometry dictionary, local mixtures correspond to tangent varieties in the same way finite mixtures give rise to secants. We use techniques from computational commutative algebra to find the local Dirac moment ideal that encodes all polynomial relations among the moments. As a corollary, we obtain the moment ideal for the Pareto distribution. For mixtures of local Diracs, an extension of Prony's method from signal processing can be used for parameter estimation. On the application side, local mixture models can be expressed as convolutions of local Diracs with distributions in an exponential family, and thus our methods can be used for these models as well.

This talk is based on joint work with Markus Wageringel.

Nonparametric Empirical Bayes Methods for Ranking with Longitudinal Data

Monday 22
14:00 - 16:00
Room 11

Jiaying Gu[1] and Roger Koenker

[1]Department of Economics, University of Toronto, Toronto, Canada
jiaying.gu@utoronto.ca

We study several compound decision problems for a large scale teacher quality evaluation using detailed administrative datasets. The longitudinal data structure permits a rich Gaussian hierarchical model with heterogeneity in both the location and scale parameters. Optimal Bayes rules are derived for effect estimation and effect ranking under various loss functions. We focus on nonparametric empirical Bayes methods which reveal some interesting features of the teacher quality distribution and allow more flexible nonlinear shrinkage rules. These results are contrasted with those obtained using the commonly used linear shrinkage method in the teacher value-added literature. In addition, one of the proposed incentive schemes to maintain accountability of the education system involves replacing teachers in the lower tail of the student performance distribution. We investigate the implementation of such policies and discuss empirical Bayes ranking methods based on the nonparametric maximum likelihood methods for general mixture models of Kiefer and Wolfowitz (1956). In particular, a close connection is revealed between the ranking problem and the multiple testing problem.

Tuesday 23
16:30 - 17:30
Second Floor

Sample Size Calculations in Simple Linear Regression: Problems and Remedies

Tianyuan Guan[1] and Marepalli Rao

[1]Biostatistics, Cincinnati, United States
guantn@mail.uc.edu

The simple linear regression is a five-parameter model involving a numeric response variable and a lone predictor. Testing the nullity of the slope parameter of the model was never a problem. The conditional test based on the least squares estimator of the slope parameter conditioned on the data of the predictor is amply adequate. However, sample size calculation for given level, power, and alternative value of the slope parameter is fraught with difficulties. The predictor is throwing a wrench in the calculations. We counter this problem by simply finding the exact distribution of the least squares estimator of the slope parameter unconditioned in every way. An alternative to the classical testing problem of the slope parameter, we could move on to testing the nullity of the correlation coefficient between the response variable and predictor. We will demonstrate that this could solve the problem and the sample size that comes out should work out for our original sample size calculation problem.

Thursday 25
10:30 - 12:30
Room 5

Towards Coding of the Asymptotic Expansion Formula in YUIMA

Emanuele Guidotti[1] and Nakahiro Yoshida

[1]Finance, University of Neuchâtel, Neuchâtel, Switzerland
emanuele.guidotti@unine.ch

Based on the Malliavin-Watanabe theory, Yoshida presented a general asymptotic expansion formula designed to facilitate implementation. The aim of this presentation is to show the current advances in the implementation of Asymptotic Expansion in YUIMA, open source academic project aimed at developing a complete environment for estimation and simulation of Stochastic Differential Equations and other Stochastic Processes via the R package called yuima and its Graphical User Interface yuimaGUI. The implementation of the asymptotic expansion formula would allow the user to efficiently compute expected values of multidimensional processes with virtually any degree of accuracy.

Financial Time Series Analysis with Unsupervised Learning

Tuesday 23
10:30 - 12:30
Room 5

Meihui Guo[1] and Ke-Jie Chen and Alan Chua and Yujung Huang

[1]Department of Applied Mathematics, National Sun Yat-Sen University, Kaohsiung, Taiwan
guomh@math.nsysu.edu.tw

We consider the prediction problem for financial time series with unknown clusters. Unsupervised learning methods, such as hierarchical and K-means clustering techniques are applied to pre-cluster time series trend. And B-spline curves are used to fit the cluster-trends. We use the conventional ARIMA models and the neural network models such as the long short-terms memory network (LSTM) models to fit the original and de-trended time series data and compare the prediction performance of these models. In the empirical study, we analyze intra-daily high frequency financial time series data.

Bayesian Inference for a Y-Linked Pedigree with Non-Syndromic Hearing Impairment

Monday 22
10:30 - 12:30
Seminari B

Cristina Gutiérrez Pérez[1] and Miguel González Velasco and Rodrigo Martínez Quintana

[1]Department of Mathematics, University of Extremadura, Badajoz, Spain
cgutierrez@unex.es

A Y-linked two-sex branching process with blind choice is a suitable stochastic model for analysing the evolution of the number of carriers of two alleles of a Y-linked gene in a two-sex monogamous population where each female chooses her partner from among the male population without caring about the allele he carries. In this work, non-parametric Bayesian inference is developed for such model assuming that the only available data are the total number of females and males (regardless of their types) up to some generation and the numbers of each type of males in the last generation. The efficiency of the methodology is illustrated with an application to a real Y-linked pedigree of a Chinese family of Tujia ethnicity with non-syndromic hearing impairment.

The research was supported by the Grant number MTM2015-70522-P (MINECO/FEDER, UE) and Grant IB16103 (Junta de Extremadura/ Fondo Europeo de Desarrollo Regional, UE).

Monday 22
16:30 - 18:30
Room 6

Trends and Trades

Olympia Hadjiliadis[1] and Michael Carlisle and Ioannis Stamos

[1]Mathematics and Statistics, Hunter College, CUNY, Brooklyn, United States
olympia.hadjiliadis@gmail.com

We construct and trading algorithm based on quickest detection considerations and study its properties.

Monday 22
16:30 - 18:30
Room 11

Clusterwise marked spatial cluster point processes, with application to single molecule microscopy

Ute Hahn[1] and Louis Gammelgaard Jensen

[1]Department of Mathematics, Aarhus University, Aarhus, Denmark
ute@math.au.dk

Spatial cluster point processes result from a superposition of finite point processes that are positioned according to an underlying parent process. In this talk, I will introduce a model of *marked* spatial cluster point processes, where the distribution of marks within a cluster depends on a mark attached to its parent. The model is motivated by single molecule microscopy: photons emitted from fluorescent molecules are registered over time, yielding space points with a time mark. Molecules start to emit photons at different points in time. They may release several photons over a period, a process called photo-blinking. In this application, the primary aim is to quantify spatial dependencies in the parent point process, namely the light emitting molecules.

I will show that the pair correlation function of the parent point process can be retrieved, up to a convolution, when the individual the mark distributions within each cluster have a relatively narrow bounded support. Our application suggests a more explicit model for the mark distributions, with marks forming a finite renewal process. The parameters of this model have direct interpretation in photo-blinking. We will discuss how to fit this model to data.

Post-Selection Inference: Risk Estimation after Data-Driven Model Selection

Tuesday 23
10:30 - 12:30
Room 10

Niels Richard Hansen[1] and Frederik Vissing Mikkelsen and Alexander Sokol

[1]Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark
Niels.R.Hansen@math.ku.dk

Valid post-selection inference procedures are statistical methods that correctly account for data-driven model selection. AIC is an estimate of risk of the MLE using the log-likelihood loss, but it is an example of an invalid risk estimate under model selection. AIC as well as other standard information criteria, Stein's unbiased risk estimate (SURE), generalized cross validation and similar statistics are biased and they systematically underestimate risk after model selection.

Biased risk estimation is due to discontinuities of the underlying parameter estimators induced by the model selection algorithms. We have derived various representations and bounds of the bias. In a special case of reduced rank regression it can be shown to vanish completely, but it is often of a considerable size.

In the talk a number of data-driven selection procedures including lasso-OLS, marginal screening, forward step-wise selection and greedy basis pursuit will be shown to fit into a novel framework for risk estimation accounting for the model selection. In this framework, the selection events are related via a flow parametrized by a tuning parameter. In addition to valid risk estimates, our framework also provides a novel leverage measure for individual observations accounting for model selection.

Statistical Learning of Ordinary Differential Equation Systems

Tuesday 23
16:30 - 17:30
Second Floor

Niels Richard Hansen[1] and Frederik Vissing Mikkelsen

[1]Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark
Niels.R.Hansen@math.ku.dk

Large systems of nonlinear ordinary differential equations (ODEs) are known to be computationally and statistically challenging to learn from data. We present adaptive integral matching (AIM) for learning polynomial or rational ODEs with a sparse network structure. The framework allows for time course data sampled from multiple environments representing e.g. different interventions or perturbations of the system. The algorithm AIM combines an initial penalised integral matching step with an adapted least squares step based on solving the ODE numerically. The R package episode implements AIM together with several other algorithms and is available from CRAN. It is shown that AIM achieves state-of-the-art network recovery for the in silico phosphoprotein abundance data from the eighth DREAM challenge with an AUROC of 0.74, and it is demonstrated via a range of numerical examples that AIM has good statistical properties while being computationally feasible even for large systems.

Monday 22
14:00 - 16:00
Seminari B

Perturbed-Variance Based Null Hypothesis Tests with an Application to Clayton Models

Wenqing He[1] and Di Shu

[1]Department of Statistical and Actuarial Sciences, University of Western Ontario, Canada
whe@stats.uwo.ca

Null hypothesis tests are popularly used when there is no appropriate alternative hypothesis available, especially in model assessment where the assumed model is evaluated with no model being considered an alternative. Motivated by the test of Clayton model in multivariate survival analysis, a simple perturbed variance resampling method is proposed for null hypothesis testing. The proposed methods make use of the perturbation method in Jin et al. (2001) to estimate the covariance matrix of the estimator to avoid possible complicated asymptotic variance estimation for the estimator. The proposed tests enjoy simplicity and a theoretical justification. We apply the proposed method to modify the tests in Shih (1998) for the assessment of Clayton models to avoid the intractable asymptotic variance. The proposed methods have simpler procedures than the parametric bootstrap and enjoy more stable performance than nonparametric bootstrap, and present promising performance as shown in the simulation studies. A colon cancer study further illustrates the proposed methods.

Thursday 25
14:00 - 16:00
Room 8

Large Domain Statistics for Brillinger-Mixing Spatial Point Processest

Lothar Heinrich[1]

[1]Institute of Mathematics, Augsburg University, Augsburg, Germany
heinrich@math.uni-augsburg.de

Large domain statistics for point processes (short:PPs) is based on a single observation of a point pattern (as realization of a stationary PP $\Psi \sim P$ on $\mathbb{R}^d$ with intensity λ) in a window W_n which grows unboundedly in all directions of $\mathbb{R}^d$ as $n \to \infty$. We assume that $\Psi \sim P$ is additionally Brillinger-mixing, that is, its reduced kth factorial cumulant measure has finite total variation for any $k \geq 2$. After repeating some basic facts on B-mixing PPs and mentioning some examples, we provide CLTs for the usual intensity estimator$\widehat{\lambda}_n$, for various empirical second-order moment functions and kernel-type product density estimators and discuss their use for testing PP hypotheses. In this connection the asymptotic variance of $\widehat{\lambda}_n$ and its consistent and asymptotically unbiased estimation play an important role. Further, we consider asymptotic goodness-of-fit tests for PPs based on scaled edge-corrected empirical (generalized) K-functions. This scaling technique allows to establish Kolmogorov-Smirnow-type tests to check PP hypotheses when intensity and K-function are known. In analogy to classical statistics two-sample tests are possible to decide whether two independent PP with known intensity and K-function have the same distribution, see L Heinrich (2018), Statistics 52(4), pp. 829-851

High-Dimensional Kendall's Tau and Spearman's Rho Correlation Matrices

Wednesday 24
10:30 - 12:30
Room 10

Johannes Heiny[1] and Thomas Mikosch and Holger Dette and Jianfeng Yao

[1]Department of Mathematics, Ruhr University Bochum, Bochum, Germany
johannes.heiny@rub.de

In many modern applications it is essential to accurately assess the dependence structure of a multivariate time series. If the dimension p and the number of observations n from this time series are of a similar order, the sample covariance matrix will typically be misleading. Aside from the high dimension, the marginal distributions of the components present another major challenge for an accurate assessment of the dependence. In the literature, one typically assumes a finite fourth moment since otherwise the largest eigenvalue of the sample covariance matrix would tend to infinity when n and p increase.

To gain insight into the dependence structure of possibly heavy-tailed time series, we provide asymptotic theory for the sample Pearson, Spearman and Kendall correlation matrices. Under heavy tails the Pearson correlation matrix R turns out to be more robust than the sample covariance matrix S. Being distribution-free measures, the sample Spearman rho correlation matrix $\mathcal{S}$ and the Kendall tau correlation matrix K capture the dependence of the underlying time series without being affected by the marginal distributions of its components. We conduct an extreme value analysis to derive the limiting distributions of the largest and smallest off-diagonal entries of $S, R, \mathcal{S}$ and K under a wide range of growth rates on the dimension p. In particular, the largest entries of all 4 matrices turn out to be Gumbel distributed. Based on this precise asymptotic result, we identify the optimal threshold levels to construct unbiased (in spectral norm) estimators of the population matrices $E[S], E[R], E[\mathcal{S}]$ and $E[K]$.

The talk is based on joint work with Thomas Mikosch (University of Copenhagen), Holger Dette (Ruhr University Bochum) and Jianfeng Yao (University of Hong Kong).

A Noise-Robust Fast Sparse Bayesian Learning Model

Monday 22
14:00 - 16:00
Room 10

Ingvild Helgoy[1] and Yushu Li

[1] Department of Mathematics, University of Bergen, Bergen, Norway
Ingvild.Helgoy@uib.no

We will propose a model which is developed within a probabilistic Bayesian supervised learning framework to detect patterns as well as carry out prediction from noisy dataset. This learning approach is fast, sparse and robust to noise. The hierarchical model structure in this Bayesian framework is designed in a way that the prior restriction will not only penalize unnecessary complexity of the model, but also depend on the variance of the random noise in the data. The hyperparameters in the model are estimated by Fast Marginal Likelihood Maximization algorithm to achieve low computational cost and faster learning process. We compare our methodology with two other Sparse Bayesian Learning models: Relevance vector machine and one Bayesian sparse model which is widely used in compressive sensing. We show that our method will generally provide more sparse solutions and is more flexible to high variance noise.

Tuesday 23
16:30 - 17:30
Second Floor

Analysis of Failure Times of a System Consisting of a Mixture of Two Components Including Inspections

Kai Hencken[1] and Markus Saltzer and Yves-Laurent Grize and Thoralf Mildenberger

[1]Corporate Research Center, ABB Switzerland Ltd, Baden-Daettwil, Switzerland
kai.hencken@ch.abb.com

In industrial applications the behavior of a decreasing failure rate at the beginning of a deployment is widely known as “infant mortality”. A variety of reasons may lead to such early failure of a product. Those can be specific assembly problems, materials problems, unintended designed flaws within requirement specification, operation outside the requirement specifications, etc. In some cases, it is not a variety of different reasons that determine the failure rate behavior at early times, but only a few reasons dominate. In this paper an analysis of the failure statistics for the case of one dominating failure reason is performed. It is discussed whether during the early phase of life for a set of components such a dominating effect can be quantified. For this an analysis of (both censored and uncensored) measured failure times are analyzed. Additional inspections, performed on a small fraction of the population with the purpose of identifying their status with respect to their belonging to one of the two populations, are done and included in the analysis. The population is modeled as a mixture of an early failing population and a group of components not failing within the foreseen lifetime. Assuming a Weibull distribution for failure times of the early failing components, the parameters of each group using a maximum likelihood approach are found. The likelihood ratio approach is used to get a confidence bound for the probability value for the early failing population. Results are given with and without the inclusion of the additional inspection. As an alternative a Bayesian analysis is done, leading to a compatible result in terms of the most probable value. Based on these results several “what-if” scenarios are discussed with the aim to give guidelines on actions how to best confirm or disprove the model and from this select the best possible repair strategy.

Thursday 25
10:30 - 12:30
Room 10

On the Application of Empirical Modelling Using Extreme Values Theory to Propose Acceptance Limits in Quality Control Processes

Raúl Hernández-Molinar[1] and Arturo Iglesias-Zárate and Roberto Aguilar-Barrón and Francisco Oviedo-Tolentino

[1]Mechanical and Electrical Department, Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico
raul.hernandez@uaslp.mx

Bootstrapping sampling processes were developed to evaluate the goodness of fit of critical quantiles that are used to calibrate special manufacturing process. The domain of the

random variables are associated to the upper end of a probability distribution function. Datasets are constructed using a random data generator system based on R. The results are verified by an automated data acquisition system; statistical tests are performed considering bias and mean square error as criteria. Results were validated showing evidence that in the case of special events occur, it is possible to use specific distributions as Gaussian, Lognormal, Weibull and Upper Extreme Value distributions. Datasets were determined, considering as acceptance limits in the upper percentiles. Analysis of the data behavior considering extreme value theory were realized. Results associated to the estimated percentiles are interpreted. It was found that based on the simulated scenarios, it is possible to conclude that certain limits can generate acceptance criteria for particular cases, with a very good performance

---

Non-Asymptotic, Universal Confidence Sets for Intrinsic Means on the Circle by Mass Concentration

Thursday 25
14:00 - 16:00
Room 6

Stefan Heyder[1] and Matthias Glock and Thomas Hotz

[1]TU Ilmenau, Ilmenau, Germany
stefan.heyder@tu-ilmenau.de

For a distribution on the circle one possible definition of its mean is given by the (set of) minimizers of expected squared distances. These so-called Fréchet means depend on the metric; choosing arc-length as the distance results in the set of intrinsic means. Given circular data which are independent and identically distributed according to this distribution, we construct confidence sets for the set of intrinsic means. Those confidence sets are guaranteed to contain all intrinsic means with a prespecified probability for any fixed sample size and any distribution; in this sense they are non-asymptotic and universal. Their construction is based on controlling both the sum of squared distances as well as its derivative by mass concentration inequalities. The advantage of this approach as opposed to confidence sets based on a central limit theorem is manifold: for one, the known central limit theorems hold only under restrictive assumptions on the distribution, in particular the intrinsic mean has to be unique; secondly, the desired coverage probability is reached only asymptotically with an unknown approximation error for the given sample size. Simulations are used to illustrate this, and to compare the proposed confidence sets with asymptotic ones.

Gamma Process Based Models for Disease Progression

Monday 22
14:00 - 16:00
Seminari B

Ayman Hijazy[1] and András Zempléni

[1]Department of probability theory and statistics, Eötvös Loránd University, Budapest, Hungary
hijaayman@gmail.com

Classic chronic diseases progression models are built by gauging the movement from the disease free state, to the preclinical (asymptomatic) one in which the disease is there but

is not manifested itself through clinical symptoms, after spending an amount of time the case then progresses to the symptomatic state.

The progression is modelled by assuming that the time spent in the disease free and the asymptomatic states are random variables following specified distributions. Estimating the parameters of these random variables leads to better planning of screening programs as well as allowing the correction of the lead time bias (apparent increase in survival observed purely due to early detection).

However, as classical approaches have shown to be sensitive to the chosen distributions and the underlying assumptions, we propose a new approach in which we model disease progression as a gamma degradation process with random starting point (onset), we derive the probabilities of cases getting detected by screens or by clinical symptoms and minimize the distance between observed and calculated distributions to get estimates of the parameters of the gamma process, screening sensitivity, sojourn time and lead time bias.

Tuesday 23
16:30 - 17:30
Second Floor

Spatio-Temporal ARMA-GARCH Model

Sondre Hølleland[1] and Hans Arnfinn Karlsen

[1]Mathematics, University of Bergen, Bergen, Norway
sondre.holleland@uib.no

We expand the ARMA-GARCH time series model to a gridded spatio-temporal situation, with both spatial and temporal dependence. In contrast to other multivariate time series models, the spatial dependence is limited due to a user-specified translation invariant neighbourhood structure, for instance having only the closest neighbours influence a given point. This makes the model (potentially) stationary and sparse. The GARCH process contributes with conditionally heteroskedastic, uncorrelated innovations in the MA part of the ARMA model. This finite, stationary model is motivated from a projection of an unlimited version. We will introduce the models, dwell on some issues, discuss estimation and show some example(s).

Tuesday 23
14:00 - 16:00
Room 8

Modeling Discrete Trade Durations with Excessive Zeros

Vladimír Holý[1] and Francisco Blasques and Petra Tomanová

[1]Department of Econometrics, University of Economics, Prague, Czech Republic
vladimir.holy@vse.cz

In finance, random times between successive transactions known as trade durations contain information about trading intensity. Usually, they are modeled by the autoregressive conditional duration (ACD) model based on a continuous distribution omitting frequent zero values. Zero durations can be caused by so-called split transactions, i.e. large orders divided into several simultaneous transactions. However, they can also be just independent transactions occurring at the same time. We propose a discrete model allowing for excessive

zero values corresponding to split transactions based on the zero-inflated negative binomial distribution. For the dynamics of time-varying mean, we utilize the generalized autoregressive score (GAS) model. We derive asymptotic properties of the maximum likelihood estimator for the proposed model. In an empirical study, we analyze 30 stocks forming Dow Jones Industrial Average (DJIA) index and find that split transactions cause on average 63

Optimal Investment for an Insurer under Levy Process with MVC Criterion

Thursday 25
10:30 - 12:30
Room 8

Leila Hosseini[1] and Navideh Modarresi

[1]Faculty of mathematics science and computer, department of mathematics, Allameh Tabataba'i University, Tehran, Iran
leila.ho3einy@gmail.com

For the cash flow of an insurance company, insurer invests into a portfolio consisting of risky and riskless assets. We consider optimal investment and risk control problems for an insurer under mean-variance-conditional value at risk criterion. The insurer's risk process is modeled by a Lévy process and it can regulate the risk by controlling the number of insurance policies. Moreover, insurers can invest in financial market consists of one risk-free asset and one risky asset whose price is described by another Lévy process. We allow the correlation among the risky asset price and risk control process. Finally, the explicit expressions for the efficient strategy and efficient frontier derived by martingale approach and then sensitivity analysis are presented to illustrate the effect of parameters on the optimal strategy.

Analysing Markov Chains using Random Measures

Thursday 25
10:30 - 12:30
Room 8

Thomas Hotz[1] and Armin Zimmermann

[1]TU Ilmenau, Ilmenau, Germany
thomas.hotz@tu-ilmenau.de

Let $X_t, t \in \mathbf{N}_0$ be a time-discrete, irreducible, positively recurrent and aperiodic Markov chain on a countable state space S with invariant distribution μ. We are interested in determining $\alpha = \int f \,\mathrm{d}\mu$ for given $f \in \mathcal{L}^2(S, \mu)$. By the ergodic theorem, $F_T = \frac{1}{T}\sum_{t=1}^{T} \to \alpha$ in quadratic mean for $T \to \infty$ if $X_0 \sim \mu$; this allows to determine α using a single, long simulation run of the Markov chain.

We generalise this approach by considering a sequence $\gamma_t, t \in \mathbf{N}_0$ of probability measures on S, computing $G_T = \frac{1}{T}\sum_{t=1}^{T} \int f \,\mathrm{d}\gamma_t$. In the talk, we give conditions which guarantee $\mathbf{E}\, F_T = \mathbf{E}\, G_T = \alpha$ while $\mathbf{Var}\, G_T \leq \mathbf{Var}\, F_T$, i.e. G_T also converges to α in quadratic mean, never slower but usually much faster than F_T.

The new measure-based approach provides much freedom to tailor the simulation to the problem at hand, e.g. when considering rare events. In fact, it can be seen as bridging the

gap between two extreme special cases: simulating a single run (or multiple runs) of the Markov chain X as well as a deterministic, numerical approach which is often infeasible as it requires unbounded memory. Simulations of concrete examples demonstrate the superiority of the measure-based approach over these extremes as well as its versatility.

Monday 22
16:30 - 18:30
Room 10

Forecast-Emphasized Principal Component for Spatial Temporal Data

Nan-Jung Hsu[1] and Han-Yueh Lee

[1]Institute of Statistics, National Tsing-Hua University, Hsin-Chu, Taiwan
njhsu@stat.nthu.edu.tw

In many applications in environmental science, it is of interest to identify dominant spatial patterns based on spatial data with temporal replicates. For this purpose, the most popular method is the empirical orthogonal function approach which is essentially the principal component analysis (PCA) for spatial data. Wang and Huang (2017) proposed a regularized PCA for handling irregularly-spaced spatial data which imposes both smoothness and sparseness constraints on the eigen-images. Following Wang and Huang's approach, this work extends the regularized PCA to identify dominant spatial-temporal patterns for spatial temporal data, in particular the temporal dynamics of spatial patterns are taking into considerations. A gradient descent method is adopted to solve the regularized spatial-temporal PCA. The effectiveness of the proposed method is investigated via simulation study and applications. It turns out that the dynamic spatial-random effect model based on the eigen-images solved in our approach has a better performance on temporal forecasts, compared to the same model based on the alternative basis from the regularized spatial PCA approach.

Monday 22
16:30 - 18:30
Room 10

False Discovery Rates to Detect Signals from Incomplete Spatially Aggregated Data

Hsin-Cheng Huang[1] and Noel Cressie and Andrew Zammit-Mangion and Guowen Huang

[1]Institute of Statistical Science, Academia Sinica, Taiepi, Taiwan
hchuang@stat.sinica.edu.tw

We consider testing for the absence/presence of a spatial signal, where the data are defined on (irregular) small areas resulting from aggregation of a continuously indexed spatial process. In socio-demographic studies, aggregation into irregular small areas or regions is often necessary to preserve confidentiality. In image analysis, the spatial domain may be discretized into pixels of different resolutions, for reasons of computational efficiency or data availability, and in remote sensing there is typically uneven coverage of Earth's surface. When all pixels at the finest resolution are observed, there are a number of ways to test for the presence of a spatial signal in the image; the one we shall focus on is a powerful nonparametric hypothesis-testing approach called the enhanced false discovery rate

(EFDR) procedure. However, the methodology relies on data defined on a rectangular spatial domain with regular pixels (small areas). In some applications, only partial information of the image is available through incomplete observation and/or the combination of some pixels into a coarser resolution. In this research, we generalize the existing methodology so that all such irregularities in the image can be handled; our idea is to augment EFDR with a Monte Carlo method in spatial statistics known as conditional simulation. We condition on the data available and simulate the full image at its finest resolution many times (M, say). Then EFDR is applied to each of these simulations to obtain M p-values for testing the null hypothesis of no signal. We test the null hypothesis of no signal in the original, irregular image by combining these statistically dependent p-values and, in the case of its rejection, we give an estimate of the signal at the finest resolution. A simulation study and an application to temperature change in the Asia-Pacific are given to demonstrate the effectiveness of the proposed procedure.

Classification of Temporal Data Using Dynamic Time Warping and Compressed Learning

Tuesday 23
10:30 - 12:30
Room 5

Shih-Feng Huang[1]

[1]National University of Kaohsiung, Kaohsiung, Taiwan
huangsf@nuk.edu.tw

This study proposes an algorithm combining the dynamic time warping (DTW) and compressed learning (CL) techniques for temporal data classification. The DTW is used to address nonsynchronous effects in multiple temporal data for determining an adequate reference trajectory. The CL is employed to represent the temporal data effectively and classify the data efficiently by cooperating with the reference trajectory. By applying the proposed algorithm and four other classification methods to several data sets, the proposed algorithm is shown to have satisfactory classification accuracies within a reasonable time. According to this advantage, the proposed algorithm is extended to establish an online monitoring system to detect different types of cardiac arrhythmia. The numerical results indicate that the online system is capable of obtaining accurate recognition results.

Multivariate Spatial Models for Severe Storm Activities

Monday 22
16:30 - 18:30
Room 10

Yen-Ning Huang[1] and Montserrat Fuentes and Eric Gilleland

[1]Statistics, Indiana University, Bloomington, United States
yenning0619@gmail.com

Extreme weather phenomena are of paramount interest for nearly every facet of society: highway safety, flood prevention measures, crop yield, general public safety, etc. As the climate changes, it is important to understand how severe weather changes. We investigate spatial patterns in two large scale indicators of severe storm environments, i.e., convective available potential energy (CAPE) and vertical wind shear (Shear), and construct a multivariate spatial model to describe the occurrence for different categories of severe storms.

Monday 22
16:30 - 18:30
Room 8

Gaussian Quasi-Maximum Likelihood Estimation for Linear GARCH Models

Feiqing Huang[1] and Guodong Li

[1]Department of Statistics & Actuarial Sciences, University of Hong Kong, Hong Kong, China
u3514252@connect.hku.hk

This paper introduces a new Filtered Historical Simulation (FHS) method based on linear GARCH for forecasting Value-at-Risk (VaR), and hence, enhances the performance when the data is heavy-tailed. Asymptotic properties including the consistency and asymptotic normality of Quasi-Maximum Likelihood Estimator (QMLE) are established. Simulation experiments are conducted to examine the finite sample performance of the method and empirical studies are also provided.

Tuesday 23
16:30 - 17:30
Second Floor

Trends Modelling in Fire Weather Index Monthly Maxima

Zuzana Hubnerova[1] and Sylvia Esterby and Steve Taylor

[1]Institute of Mathematics, Brno University of Technology, Brno, Czech Republic
zuzana.hubnerova@gmail.com

The Fire Weather Index (FWI), a numeric rating of wildland fire intensity, is one of the three fire behaviour indices of the Canadian Forest Fire Weather Index System. Calculation of the components of the system is based on consecutive daily observations of temperature, relative humidity, wind speed, and 24-hour rainfall. That is why a potential climate change might be expected to influence the values of FWI throughout time. Our poster presents the results of an analysis of FWI records within the years 1970 to 2014 at several stations in British Columbia, Canada.

Because of a typical variation of the FWI between the months, the monthly FWI maxima were of interest. Both univariate and spatial (max-stable) models with generalized extreme-value distribution with possible linear dependence on time in both location and scale parameters were applied to assess a temporal dependence of the monthly FWI maxima.

Predicting LGD Distributions with Mixed Continuous and Discrete Ordinal Outcomes

Monday 22
10:30 - 12:30
Room 12

Ruey-Ching Hwang[1] and Chih-Kang Chu

[1]Department of Finance, National Dong Hwa University, Hualien, Taiwan
rchwang@gms.ndhu.edu.tw

For predicting the loss given default (LGD) distribution, due to its complex structure, the ordinary modeling approach using the parametric density function may suffer a loss of predictive power. To overcome this potential drawback, we use the cumulative probability model (CPM) to predict the LGD distribution. The CPM applies a transformed variable to model the LGD distribution. The transformed variable has a semiparametric structure. It models the predictor effects parametrically. The functional form of the transformation is unspecified. To implement CPM, we design a simple procedure based on the logistic regression to estimate its parameters and collect a sample of defaulted debts from Moody's Default and Recovery Database. Given this sample, we use an expanding rolling window approach to investigate the out-of-time performance of CPM and its alternatives. Our results confirm that CPM is better than its alternatives, in the sense of yielding more accurate LGD distribution predictions.

Right-Censored Mixed Poisson Count Models with Detection Times

Friday 26
09:00 - 11:00
Room 5

Wen-Han Hwang[1] and Jakub Stoklosa

[1]Institute of Statistics, National Chung Hsing University, Taichung, Taiwan
wenhan@nchu.edu.tw

Conducting complete surveys on flora and fauna species within a given area of interest can be costly, particularly if there is an over-abundance of several species. A commonly used approach, which aims to reduce time and costs, consists of censoring data during the collecting process until some set threshold is reached – e.g., rather than counting every individual, the survey is stopped as soon as one or two counts have been observed. Although this approach is cheaper to conduct than a complete survey, some statistical efficiency in model estimators is lost. We propose a new class of regression estimation models for count data that incorporate information from detection times (or catch effort) that are collected during sampling. First, we show that incorporating ancillary information in the form of detection times can greatly improve statistical efficiency over, say, right-censored Poisson or negative binomial models. Furthermore, the proposed models retain the same cost-effectiveness as censored-type models. These models can be extended to a more general class of mixed Poisson models. We investigate model performance on simulated data and give several examples using real data.

Tuesday 23
10:30 - 12:30
Room 7

A Method for Constructing a Circular Distribution from a Single Valued Analytic Function

Tomoaki Imoto[1]

[1]School of Management and Information, University of Shizuoka, Shizuoka, Japan
imoto@u-shizuoka-ken.ac.jp

In many diverse scientific fields, there often appear data considered as points on a unit circle, called circular data, such as wind directions at a monitoring site, vanishing angles of birds from a starting point, intensive care unit arrival times on the 24-hour clock. For modelling circular data, many continuous circular distributions have been constructed using several methods, i.e., projection method, conditioning method, maximizing entropy method, wrapping method and so on. In my talk, we consider the method for constructing a circular distribution by a based single-valued analytic function. The constructed distribution has simple expressions for probability density and characteristic functions, and the parameters in it play a role of controlling its first and second trigonometric moments. This means that, if the distribution is unimodal, the parameters measure its location, concentration, skewness and kurtosis. The properties like unimodality and the point of mode can be found through the curve on the complex plane drawn by a based single-valued analytic function. Since the probability density function and its characteristic function are expressed by simple forms, we can straightforwardly estimate the parameters by method of moments and estimation based on the characteristic function.

Tuesday 23
16:30 - 17:30
Second Floor

Local Linear Estimators in Quantile Regression with Errors-in-Variables

Dimitris Ioannidis[1] and Dimitris Ioannides and Eric Matzner

[1]University of Macedonia, Thessaloniki, Greece
dimioan@uom.edu.gr

Dimitris. A. Ioannides, and E. Matzner-Lober 2

1Department of Economics, University of Macedonia, Greece (email: dimioan@uom.edu.gr)

2Department of Statistics, University of Rennes, France (email: eml@uhb.fr)

Local linear quantile regression are popular techniques for nonparametric estimation and have received great attention in the literature for data without measurement errors. Quantile regression is useful for constructing predictive intervals , estimating reference curves or detecting heteroscedasticity. However, in many practical situations the explanatory variables are measured with error, or are observed through a proxy. In this case the local constant quantile estimator, can be easily estimated by exploiting its similarity with the deconvolution of kernel density estimator . A local linear quantile estimator r in the errors-in-variables context is proposed, and its asymptotic properties are obtained. Finally, its finite sample performance is analyzed via a large simulation study .

References: Delaigle A., Fan J and Raymond C. (2009). A Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem , JASA, 104, 348-359. D. Ioannides and

E. Matzner (2009). Regression quantiles with errors-in-variables, Journal of Nonparametric in Statistics, 16, 515-524.

Variance Reduction via Empirical Variance Minimization

Tuesday 23
14:00 - 16:00
Seminari B

Leonid Iosipoi[1] and Denis Belomestny and Nikita Zhivotovskiy

[1]Computer Science, National Research University Higher School of Economics, Moscow
liosipoi@hse.ru

We propose and study a generic variance reduction approach. The proposed method is based on minimization of the empirical variance over a suitable class of zero mean control functionals. We discuss several possibilities of construct- ing zero mean control functionals and present non-asymptotic error bounds for the variance reduced Monte Carlo estimates. Finally, a simulation study showing nu- merical efficiency of the proposed approach is presented.

Automatic Text Simplification

Thursday 25
14:00 - 16:00
Room 6

El Mehdi Issouani[1] and Patrice Bertail and Mélanie Zetlaoui and Thierry Dumont

[1]Paris Nanterre University, Nanterre, France
emissouani@gmail.com

We first introduce methods that can be used to generate classifiers for Part-of-Speech tagging problems utilizing feature-based models. The key to the proposed technique is the combination of a "feature" extraction stage composing the data in an « elementary space » to produce "features", and a classification stage utilizing a Maximum Entropy Principle to estimate the model (ie. to learn the weight of each "feature" extracted). In a second time, we'll use the same approach to construct a classifier which gets a sentence and then returns an output that indicates whether the input sentence is complex or simple.

A New Central Limit Theorem for GARCH Processes without Kesten's Regularity

Thursday 25
10:30 - 12:30
Room 10

Adam Jakubowski[1] and Zbigniew S. Szewczak

[1]Department of Probability and Stochastic Analysis, Nicolaus Copernicus University, Toruń, Poland
adjakubo@mat.umk.pl

We consider a class of perpetuities which admit direct characterization of asymptotics of the key truncated moment. The class contains perpetuities without polynomial decay of tail probabilities and thus not satisfying Kesten's theorem. We show how to apply the results in deriving a new central limit theorem for GARCH(1,1) processes in the critical case.

Thursday 25
14:00 - 16:00
Room 10

Scoring Predictions at Extreme Quantiles

Kaushik Jana[1] and Axel Gandy and Almut Veraart

[1]The Alan Turing Institute, Imperial College London, London, United Kingdom
k.jana@imperial.ac.uk

Prediction of extreme quantiles lying outside the range of historical data is of interest in numerous applications. Extreme value theory provides various competing models through different assumptions on tail behaviors of the underlying distribution. Assessment of competing estimates based on their predictive performance is a useful way to select the best estimate in a given situation. However, due to extreme nature of this inference problem, it might be very well possible that, the predicted quantiles are not seen in the historical records. Therefore, making it challenging to validate the prediction with its realization. In this article, we propose a non-parametric scoring method for high quantile estimates by predicting a sequence of equally extremal quantiles on different parts of the data. We then use cross validation approach and covariate information for scoring the competing estimates. The performance of the scoring method is assessed in a simulation study that mimics the practical application at hand. This methodology is applied to motivating netflow data available from the Los Alamos National Laboratory.

Monday 22
14:00 - 16:00
Room 8

Multiscale Local Polynomial Estimation from Highly Irregular Data

Maarten Jansen[1]

[1]Mathematics, Université libre de Bruxelles, Brussels, Belgium
maarten.jansen@ulb.ac.be

The multiscale local polynomial transform (MLPT) combines features from kernel and local polynomial smoothing with sparse, multiscale analysis from wavelet transforms. Like Deslauriers-Dubuc (DD) or B-spline wavelet transforms, the MLPT can be applied to data on irregular grids of covariates. The MLPT shares with the DD wavelets the property that observations can be used without preprocessing. An important benefit from a MLPT analysis is the multiscale bandwidth, which acts as a user controlled definition of the scales in a multiscale analysis. On higly irregular data, a careful, covariate dependent refinement scheme is needed in order to properly process the data at the correct scales. The analysis is finetuned for the bandwidths, especially the bandwidth at finest resolution. Further parameters include the order of local polynomial, and parameters for variance propagation control. The choice of these parameters is topic of ongoing research. Applications in scattered data smoothing and density estimation are discussed.

Extremal Markov Chains Driven by the Kendall Convolution

Thursday 25
10:30 - 12:30
Room 8

Barbara Jasiulis-Gołdyn[1] and Marek Arendarczyk and Jolanta Misiewicz and Edward Omey

[1]Mathematical Institute, University of Wrocław, Wrocław, Poland
jasiulis@math.uni.wroc.pl

We consider the Kendall convolution algebra such that the result of two deltas is the Pareto probability measure. The main goals of this talk are focused on the renewal theory , some aspects of the fluctuation theory and asymptotic properties for the extremal Markov sequences of the Kendall type (called also Kendall random walks), which are discrete time Levy processes (see [2]) under the Kendall convolution.

We show that one dimensional distributions of Kendall random walks are regularly varying and prove finite-dimensional convergence for continuous time stochastic processes based on Kendall random walks. We construct renewal processes based on the extremal Markov sequences and investigate limit behaviour of the corresponding renewal function (for details see [3]). Limit Theorems for Kendall random walks and corresponding renewal processes [1,3] will be showed using regular variation and Williamsom transform.

This work is a part of project "First order Kendall maximal autoregressive processes and their applications", which is carried out within the POWROTY/REINTEGRATION programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.

[1] M. Arendarczyk, B.H. Jasiulis-Gołdyn, E.A.M. Omey, Asymptotic properties of extremal Markov chains of Kendall type, submitted, https://arxiv.org/pdf/1901.05698.pdf, 2019.

[2] M. Borowiecka-Olszewska, B.H. Jasiulis-Gołdyn, J.K. Misiewicz, J. Rosiński, Lévy processes and stochastic integral in the sense of generalized convolution, Bernoulli, 21(4), 2513-2551, 2015.

[3] B.H. Jasiulis-Gołdyn, K. Naskręt, J.K. Misiewicz, E.A.M. Omey, Renewal theory for extremal Markov sequences of the Kendall type, submitted, arXiv: https://arxiv.org/pdf/1803.11090.pdf, 2019.

Fourth Cumulant for Multivariate Aggregate Claim Models

Wednesday 24
10:30 - 12:30
Room 6

Farrukh Javed[1] and Stepan Mazur and Nicola Loperfido

[1]Department of Statistics, Örebro University, Örebro, Sweden
Farrukh.Javed@oru.se

The fourth cumulant for the aggregated multivariate claims is considered. A formula is presented for the general case when the aggregating variable is independent of the multivariate claims. Two important special cases are considered. In the first one, multivariate skewed normal claims are considered and aggregated by a Poisson variable. The second

case is dealing with multivariate asymmetric generalized Laplace and aggregation is made by a negative binomial variable. Due to the invariance property the latter case can be derived directly, leading to the identity involving the cumulant of the claims and the aggregated claims. There is a well established relation between asymmetric Laplace motion and negative binomial process that corresponds to the invariance principle of the aggregating claims for the generalized asymmetric Laplace distribution. We explore this relation and provide multivariate continuous time version of the results. It is discussed how these results that deals only with dependence in the claim sizes can be used to obtain a formula for the fourth cumulant for more complex aggregate models of multivariate claims in which the dependence is also in the aggregating variables.

Tuesday 23
14:00 - 16:00
Room 7

Bootstrapping Hill Estimator and Tail Arrays Sums for Regularly Varying Time Series

Carsten Jentsch[1] and Rafał Kulik

[1]Faculty of Statistics, TU Dortmund University, Dortmund, Germany
jentsch@statistik.tu-dortmund.de

In the extreme value analysis of stationary regularly varying time series, tail array sums form a broad class of statistics suitable to analyze their extremal behavior. Asymptotic theory has been developed by Rootzén et al. (1998) under mixing conditions and in Kulik et al. (2015) for functions of geometrically ergodic Markov chains. A more general framework of cluster functionals is presented in Drees and Rootzén (2010). The resulting limiting distributions are complex and cumbersome to estimate such that suitable bootstrap procedures are desired, but available bootstrap consistency results are scarce.

We propose a multiplier block bootstrap to estimate the distribution of tail array sums. We prove that, conditionally on the data, an appropriately constructed multiplier block bootstrap statistic does converge to the correct limiting distribution. In contrast, an apparently natural, but naive application of the multiplier block bootstrap scheme does not yield the correct limit.

We provide numerical evidence of our theoretical findings and illustrate the superiority of the proposed multiplier block bootstrap over some obvious competitors. Moreover, it proves to be computationally more efficient.

Friday 26
09:00 - 11:00
Room 6

Additive Regression with Mixed Predictors and Incomplete Hilbertian Responses

Jeongmin Jeon[1] and Byeong U. Park and Ingrid Van Keilegom

[1]Research Centre for Operational Research and Business Statistics, KU Leuven, Leuven, Belgium
jeongmin.jeon@kuleuven.be

In this talk, we consider a regression problem when there are mixed continuous and discrete Euclidean predictors and incompletely observed Hilbertian responses. The Hilbertian responses cover not only functional responses and Euclidean responses but also density-valued responses and simplex-valued responses. We also cover the case that the response variable is real-valued and subject to censoring or missing. Meanwhile, when there are mixed continuous and discrete Euclidean predictors, commonly adopted nonparametric structured models are partially linear models and varying coefficient models. Here, we consider a fully nonparametric additive model fitted by a novel approach based on the smooth backfitting method. We establish a complete asymptotic theory including rates of convergence and asymptotic distributions. Our simulations and real data analysis show that the proposed method is valid for various types of responses and predictors.

A Multidimensional Latent Variable Analysis for Longitudinal Data

Friday 26
09:00 - 11:00
Room 5

Saebom Jeon[1]

[1]Mokwon, Daejeon, Korea, South
alwaysns@mokwon.ac.kr

This article proposes a joint latent class analysis for longitudinal data, joint latent class profile analysis (JLPCA), which provides a principle for the systematic identification of the sequential patterns for multiple discrete latent variables. Inferences of parameters are obtained by an expectation-maximization algorithm using recursive formula. We apply JLCPA using data from the 'National Longitudinal Survey of Youth 1997 (NLSY97)', to investigate progressions as well as identifications of adolescent drug-using behaviors; alcohol, cigarette, and marijuana use across time. As early-onset drinking is a well-known risk factor for adverse public health consequences, such as abuse of multiple substances, we focused on the multiple drug-taking behavior and its stage-sequential pattern for early-onset drinkers. The JLCPA uncovers four profiles of simultaneous patterns of substance use representing the most probable sequences of drug-use behaviors in youth.

A New Graphical Device and Related Tests for the Shape of Non-Parametric Regression Function

Monday 22
14:00 - 16:00
Room 8

Prashant Jha[1] and Subhra Sankar Dhar and Mohammad Arshad Rahman and Joydeep Dutta

[1]Department of Mathematics and Statistics, IIT Kanpur, Kanpur, India
pjha@iitk.ac.in

We consider a non-parametric regression model $y = m(x)+\epsilon$, and propose a novel graphical device to check whether the v-th derivative of the regression function $m(x)$ is positive or otherwise. Since the shape of the regression function can be completely characterized by its derivatives, the graphical device can correctly identify the shape of the regression function. The proposed device includes the check for monotonicity and convexity of the function as

special cases. We also present an example to elucidate the practical utility of the graphical device. In addition, we employ the graphical device to formulate a class of test statistics and derive its asymptotic distribution. The tests are exhibited in various simulated and real data examples.

Tuesday 23
16:30 - 17:30
Second Floor

Bootstrap of Nonparametric Dynamic Discrete Choice Model

Seongjae Joo[1]

[1]Department of Statistics, Seoul National University, Seoul, Korea, South
jsjae2000@gmail.com

We investigate model based bootstrap for nonparametric dynamic discrete response models. Data encompass discrete response variable and other covariates which may include either continuous or discrete variables. Moreover, covariates contain lags of response variable. We focus on local linear estimator maximizing local likelihood, which have consistency and asymptotic normality under some conditions. Based on this estimator, we propose two model based bootstrap procedures : (i) generate response variables given original data from regression model ignoring the dynamic structures of original data (ii) generate bootstrap samples mimicking the dynamic structure of the original process. Generating bootstrap samples, we calculate the one-step bootstrap estimators, which dramatically reduce computational burden than maximizing bootstrap local likelihood. We will show the first-order asymptotic validity of both bootstrap procedures. In simulation study, we compare empirical coverage of confidence interval based on asymptotic distribution as well as that of proposed bootstrap confidence intervals.

Wednesday 24
10:30 - 12:30
Room 11

Parallel Lines and Rotations of Crystals

Peter Jupp[1] and Ian Goudie and Richard Arnold and Helmut Schaeben

[1]School of Mathematics and Statistics, University of St Andrews, St Andrews, United Kingdom
pej@st-andrews.ac.uk

Two practical problems are described and each is translated into directional terms. Some inferential aspects are discussed. Parallel lines: In some geological contexts points in a plane are observed and interest lies in assessing whether or not they lie near some set of equally-spaced parallel straight lines. Closeness of points $x_1, \ldots, x_n$ to lines normal to the unit vector u and distance q apart is equivalent to the differences of the numbers x_i^T, u being multiples of q. Thus this problem is a 2-dimensional version of D.G. Kendall's (1974, Phil. Trans. A) 'quantum hunting' problem. A reasonable von Mises assumption leads to a likelihood that is infinitely differentiable but very bumpy.

Rotations of crystals: Orientations of asymmetrical rigid objects in R^3 correspond to elements of the rotation group, $SO(3)$. Orientations of objects (such as crystals) with symmetry group K correspond to elements of the space $SO(3)/K$ of left cosets. Embedding

$SO(3)/K$ into suitable spaces of tensors (Arnold, et al., 2018, JMVA) enables inference on $SO(3)/K$. Some metallurgical processes lead to a change of symmetry group of crystallites from K_1 to K_2. These are modelled here using regression functions of the form $[U]_1 \to [RUA]_2$, where $R, U, A \in SO(3)$ and e.g. $[U]_j$ denotes the equivalence class of U in $SO(3)/K_j, (j = 1, 2)$.

Nonparametric Tests of Symmetry for Non-Elliptical Distributions

Tuesday 23
16:30 - 17:30
Second Floor

Jan Kalina[1]

[1]The Czech Academy of Sciences, Institute of Information Theory and Automation, Praha 8, Czech Republic
kalina@cs.cas.cz

The article discusses how various multivariate location and scatter estimators capture the symmetry of the underlying distribution. Very general sufficient conditions are formulated, which ensure various symmetry properties of functionals corresponding to location or scatter. Examples of robust multivariate estimators, which fulfil these conditions, are discussed in detail. The obtained symmetry of the estimators is applicable to hypothesis tests of symmetry of the underlying distribution of the multivariate data. For this task, we propose to perform permutation tests exploiting the nonparametric combination methodology. The performance of the newly proposed tests is illustrated on simulated as well as real data. The tests are suitable for small sample sizes and represent the first available symmetry tests suitable also for non-elliptical distributions and for more than just two variables.

On Bayesian Structural Equation Modelling

Monday 22
14:00 - 16:00
Room 10

Kostas Kalogeropoulos[1] and Konstantinos Vamvourellis and Irini Moustaki

[1]Statistics London School of Economics, London, United Kingdom
k.kalogeropoulos@lse.ac.uk

Bayesian analysis can be used to provide a more flexible approach to factor analysis and structural equation modeling. In this setting non-Bayesian approaches maybe thought of as setting several parameters, such as cross factor loading and conditional covariance to zero according to the hypothesized theory. The Bayesian approach replaces these exact zeros with approximate zeros based on informative priors. This results to a more natural and flexible modelling framework but care must be taken to identify whether a model is a good fit to the data. In this work we extend the work of Muthén, B. & Asparouhov, T. (2012) by formulating at the individual level. This allows i. to identify individuals for which the model is not appropriate in addition to population parameters and ii. to cover non-Gaussian models such logistic item response theory. Moreover we explore the use of out of sample fit indices to explore model assessment and fit. The methodology is illustrated on both real and simulated data.

Thursday 25
14:00 - 16:00
Room 7

How many Filters are Enough? a New Algorithm for Eigenvector Space-Time Filtering

Yiannis Kamarianakis[1] and Meng Wang

[1]Institute of Applied & Computational Mathematics, Foundation for Research & Technology -Hellas, Heraklion Crete, Greece
kamarian@iacm.forth.gr

Moran's eigenvector filtering (EF) is a popular technique for the analysis of spatial or space-time lattice data. In spatial data analyses, EF is based on synthetic predictors which represent distinct map patterns; these control variables capture stochastic spatial dependencies in the residuals, allowing model building to proceed as if observations were independent. In this talk I will discuss how EF compares to well-known spatial econometric models and Geographically Weighted Regressions. A new algorithm for EF filtering will be presented; the algorithm combines variance-inflation-factor (VIF) filtering, correlation-based screening and a fast adaptive lasso estimator with EF. The new scheme possesses a number of advantages: it can be implemented to estimate challenging specifications (e.g. space-time models with spatially varying coefficients) and frequently it is fast enough to allow for bootstrap confidence intervals. In a series on Monte Carlo experiments, the proposed method is compared against existing Hierarchical Bayesian specifications and conventional forward selection schemes for EF. Finally, an application to the analysis of regional climate model outputs will be discussed.

Wednesday 24
10:30 - 12:30
Room 7

Scaling Limits of Piecewise Deterministic Monte Carlo Methods

Kengo Kamatani[1] and Gareth Roberts and Joris Bierkens

[1]Engineering Science, Osaka University, Osaka, Japan
kamatani@sigmath.es.osaka-u.ac.jp

Piecewise deterministic Markov processes are an important new tool in the design of Markov Chain Monte Carlo algorithms. Two examples of fundamental importance are the Bouncy Particle Sampler and the Zig-Zag process. In this talk scaling limits for both algorithms are determined. Here the dimensionality of the space tends towards infinity and the target distribution is the multivariate standard normal distribution. For several quantities of interest, the scaling limits show qualitatively very different and rich behavior. Based on these scaling limits the performance of the two algorithms in high dimensions can be compared. Finally, we provide a criterion for the choice of the refreshment rate of the bouncy particle sampler. This is a joint work with J. Bierkens (TU Delft) and G. O. Roberts (Warwick)

Nonparametric Estimation of the Kernel Function of Stable Moving Average Processes

Monday 22
16:30 - 18:30
Room 7

Jürgen Kampf[1] and Georgiy Shevchenko and Evgeny Spodarev

[1]Institute of Mathematics, University of Rostock, Rostock, Germany
juergen.kampf@uni-rostock.de

In this talk we will consider moving average random fields, i.e. random fields which are given as the convolution of an unknown, deterministic function f, which is called the kernel function, and a symmetric stable random measure. For example, CARMA processes are of this form. Our aim is to derive a non-parametric estimator for f. We assume that the points at which the random field is observed form a lattice and we consider the asymptotics as the grid size tends to zero and at the same time the area covered by the lattice tends to infinity.

Under the assumption that f is symmetric, of positive type and has L^2-norm 1, we obtain a weakly consistent estimator $\tilde{f}$ for f, where the distance between f and $\tilde{f}$ is measured in L^2-sense.

An essential tool for the construction of this estimator will be the self-normalized periodogram. We will show that a smoothed version of the self-normalized periodogram converges in probability to the normalized squared absolute value of the Fourier transform of f. Under the regularity assumptions mentioned above we can thus reconstruct the Fourier transform of f and so we are able to reconstruct f itself.

Semiparametric Quantile Regression using Quantile-Based Asymmetric Family of Densities

Monday 22
14:00 - 16:00
Room 6

Rezaul Karim[1] and Irène Gijbels and Anneleen Verhasselt

[1]Mathematics, KU Leuven, Hasselt, Belgium
mdrezaul.karim@kuleuven.be

Quantile regression is an important tool in data analysis. Linear regression, or more generally, parametric quantile regression imposes often too restrictive assumptions. Nonparametric regression avoids making distributional assumptions, but might have the disadvantage of not exploiting distributional modeling elements that might be brought in. In this talk, a semiparametric approach towards estimating conditional quantile curves will be discussed. It is based on a recently studied large family of asymmetric densities of which the location parameter is a quantile (and not a mean). Passing to conditional densities and exploiting local likelihood techniques in a multiparameter functional setting then leads to a semiparametric estimation procedure. For the local maximum likelihood estimators we establish asymptotic distributional properties. The appealing semiparametric framework allows to study in detail the bandwidth selection issue, and provide several practical bandwidth selectors. The practical use of the semiparametric method is illustrated in the analysis of maximum winds speeds of hurricanes in the North Atlantic region. A small simulation study includes a comparison with nonparametric local linear quantile regression as well as a brief investigation of robustness against miss-specifying the parametric model part.

Monday 22
16:30 - 18:30
Room 10

A Spatial ARMA-GARCH Model

Hans Arnfinn Karlsen[1] and Sondre Hølleland and Sondre Hølleland

[1]Department of Mathematics, University of Bergen, Hjellestad, Norway
hans.karlsen@uib.no

We consider a spatial ARMA-GARCH model on an infinite spatial grid. In particular this model contains a spatial temporal extension of the standard univariate GARCH model. In this work we find operational conditions for the existence of a unique stationary ergodic solution of the model. Moreover, the GARCH part implies an extended ACF for the squared variables. Important properties of this ACF depend on the local stationary neighborhood network implied by the model. For estimation of the unknown parameter vector a conditional likelihood method, as used in the time series case, suffers from a boundary that increases with sample size. However, the problem is here handled by maximizing a modified likelihood function. This gives both consistency and asymptotic normality. The spatial temporal version of the model is also related to its projection onto a stationary model living on a finite spatial region but with unrestricted time dimension.

Tuesday 23
14:00 - 16:00
Seminari B

An Investigation of the Performances of Liu Estimators in a Beta Regression Context in the Presence of Multicollinearity.

Peter Karlsson[1] and Pär Sjölander and Kristofer Månsson and B.M. Golam Kibria

[1]Department of Economics and Statistics, Linnaeus University, Växjö, Sweden
peter.s.karlson@lnu.se

Beta regression has become a popular tool for perform a regression analysis in which the dependent variable is restricted to the interval from 0 to 1, such as rates and proportions. In this paper we suggest a new Liu type of estimator for this model that may be used in a number of realistic situations when different degree of correlations among the regressors is present. By using Monte Carlo methods, we investigate the properties of five different estimators of the shrinkage parameter and compare it with the maximum likelihood estimator (MLE). The results from this investigation deceptive indicate that in the presence of multicollinearity the Liu type estimator outperforms the MLE. Surprisingly if variability of the dependent variable is high then the it is found that the Liu type of estimator outperforms MLE even when the regressor are strict orthogonal. Finally in an empirical application we show the benefit of the new approach.

Multivariate Associations Measures

Friday 26
09:00 - 11:00
Room 8

Vojtěch Kika[1] and Irène Gijbels and Marek Omelka

[1]Department of Mathematics, KU Leuven, Leuven, Belgium
vojtech.kika@kuleuven.be

Coefficients like Kendall's tau or Spearman's rho, used to measure (strength of) an association between two random variables, were thoroughly studied and described in the middle of 20th century. Requirements on bivariate association measures are well-known. However, generalization of such measures into higher dimensions is not very straightforward and brings discussion on the desirable properties. In addition, bivariate measures can be often generalized in multiple manners. We discuss methods of generalization, their theoretical properties, strengths and weaknesses. In particular, behavior of association measures when dimension grows to infinity is of interest. Moreover, examples illustrate the performance and use of such measures.

Nonparametric Regression Kriging with Varying Coefficient Regression Models

Tuesday 23
16:30 - 17:30
Second Floor

Min Jung Kim[1] and Chae Young Lim and Byeong U. Park

[1]Department of Statistics, Seoul National University, Seoul, Korea, South
mjkim.stat@gmail.com

Varying coefficient models are very useful for explaining complex relationships between a response and covariates. In this paper, we propose a spatial prediction methodology that combines a varying coefficient regression with kriging of the regression residuals. To estimate the coefficient functions, we employ a smooth backfitting algorithm. We establish the convergence of algorithm and derive the asymptotic normality of the estimators for alpha-mixing spatial process. The proposed method is illustrated by numerical studies.

Revisiting Minimax Property

Thursday 25
14:00 - 16:00
Room 8

Alisa Kirichenko[1] and Peter Grunwald

[1]CWI, Amsterdam, Netherlands
alice.kirichenko@gmail.com

In model selection it is desirable to have a statistical method that will be, on one hand, consistent in choosing the right model and on the other hand, perform optimally in post-selection estimation. Yang (2005) shows that achieving both at the same time is impossible: any consistent method performs sub-optimally in the minimax sense. In the light of this result, it is not surprising that the popular AIC, although being minimax optimal, is inconsistent. Contrarily, BIC is consistent in choosing the small model containing the truth

for large enough sample sizes, but misses the minimax rate by a logarithmic factor. In this talk I revisit the definition of the minimax property in order to develop a new, more suitable metric that allows apt comparison of existing model selection methods. I show how one can benefit from using martingale theory for deriving robust minimax rates.

Tuesday 23
14:00 - 16:00
Seminari B

Asymptotic Equivalence of Inference Post-Model-Selection with the Lasso on Randomized Data and on a Subset of the Data

Danijel Kivaranovic[1] and Hannes Leeb

[1]Department of Statistics and Operations Research, University of Vienna, Vienna, Austria
danijel.kivaranovic@univie.ac.at

Inference post-model-selection on randomized data (Tian & Taylor (2018)) and post-model-selection on a subset of the data (dubbed "data carving" by Fithian et al. (2017)) were introduced as more powerful alternatives to the polyhedral method of Lee et al. (2016), which is known to produce confidence intervals of diverging length as shown by Kivaranovic & Leeb (2018). We show that both inference procedures are asymptotically equivalent in a Gaussian regression setting where the lasso is used as model selector. The key to this equivalence is a new multivariate CLT-type result. If the randomization is sufficiently strong in case of model selection on randomized data, or if the subset is sufficiently small in case of model selection on a subset of the data, our result implies total variation convergence of the post-model-selection estimator to a Gaussian distribution. In terms of confidence interval length, our result implies that the length of confidence intervals with exact conditional coverage post-model-selection on randomized data or on a subset of the data converges to the length of unconditional Gaussian confidence intervals if the randomization is sufficiently strong, or if the subset is sufficiently small, respectively.

Monday 22
14:00 - 16:00
Room 8

Two Goodness of Fit Tests in Nonparametric Transformation Models

Nick Kloodt[1] and Natalie Neumeyer and Ingrid Van Keilegom

[1]Mathematical Statistics and Stochastic Processes, University of Hamburg, Hamburg, Germany
nick.kloodt@uni-hamburg.de

Over the last years, transformation models have attracted more and more attention since they are often used to obtain desirable properties by first transforming the dependent random variable of a regression model. Applications for such models can reach from reducing skewness of the data to inducing additivity, homoscedasticity or even normality of the error terms. While completely nonparametric modeling is possible this might be accompanied by a loss in terms of estimating performance. Hence, it is desirable to test for the type of the model a priori in order to apply appropriate estimating approaches afterwards. In this talk, two goodness of fit tests are presented, which on the one hand test for a parametric

transformation function and on the other hand for a parametric regression function in the setting of a nonparametric transformation model. Asymptotic results about convergence rates of the test statistics as well as consistency of the corresponding tests, assosiated with a small simulation study, are given.

Estimation of Causal CARMA Random Fields

Monday 22
16:30 - 18:30
Room 7

Claudia Klueppelberg[1] and Viet Son Pham

[1]Department of Mathematics, Technical University of Munich, Germany
cklu@ma.tum.de

We estimate model parameters of Lévy-driven causal CARMA random fields by a semi-parametric method. In a first step, we estimate the empirical variogram and, in a second step, fit it to its theoretical parametric counterpart by a weighted least squares (WLS) approach. Based on asymptotic results for the empirical variogram, we show strong consistency and asymptotic normality of the parameter estimator. Furthermore, we conduct a simulation study to assess the quality of the WLS estimator for finite samples. For the simulation we utilize numerical approximation schemes based on truncation and discretization of stochastic integrals and we analyze the associated simulation errors in detail. Finally, we apply our results to real data of the cosmic microwave background.

This is joint work with Viet Son Pham.

Statistical Inference and Data Analysis on Length Metric Spaces

Tuesday 23
10:30 - 12:30
Room 7

Kei Kobayashi[1] and Henry Wynn

[1]Mathematics, Keio University, Yokohama, Japan
kei@math.keio.ac.jp

A methodology is developed for data analysis based on length spaces (geodesic metric spaces). Observed data is assumed to be distributed on/around a length space and the metric should be estimated if it is unknown. Such metrics are transformed in a number of ways to produce parametrised families of length spaces, empirical versions of which allow computation of intrinsic means and associated measures of dispersion. These reveal properties of the data, based on geometry, such as those that are difficult to see from the raw Euclidean distances. Examples of application include clustering and classification. We reveal how curvature of the length space plays important roles in such data analysis. Because our study considers length spaces rather than Riemannian manifolds, it is possible to argue their curvature even for non-smooth spaces such as empirical graphs and simplicial complexes. We propose a method to change the length space to be CAT(0) and the intrinsic means to be unique. Moreover, a so-called "metric cone" construction is proposed to argue the curvature even for non-geodesic metric spaces. It is shown how to empirically tune the curvature of the metric spaces, making it possible to apply them to a number of real cases.

Thursday 25
10:30 - 12:30
Room 5

High-Dimensional Covariance Estimation in YUIMA Package

Yuta Koike[1]

[1]Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, Japan
kyuta@ms.u-tokyo.ac.jp

Covariance matrix estimation is one of the most actively studied topics in high-frequency financial econometrics. Recently, the high-dimensionality problem has attracted much attention in this research area. Although the "cce" function contained in the R package "yuima" can flexibly estimate the covariance matrix of high-frequency data with a simple command, it does not cover the recent progress in high-dimensional covariance matrix estimation from high-frequency data. This talk presents our attempt to fill in this gap.

Monday 22
16:30 - 18:30
Room 10

Spatially Explicit Capture Recapture as a Self Exciting Point Process

Aleksandar Kolev[1] and Gordon Ross and Ruth King and David Borchers

[1]Department of Statistical Science, University College London, London, United Kingdom
kolev.al.at@gmail.com

Hawkes process is a self-exciting point process which provides the opportunity for multi-layered intensity structure that addresses the rate of events as a function of previous events' history. Triggering and clustering behaviours are naturally captured. A generalization of this model is the Epidemic-Type Aftershock Sequence (ETAS) which is one of the best-performing methods for seismic forecasting. We extended the most recent development of the spatial-ETAS model to address the concept of animal movement as a point process. Spatially explicit capture recapture (SECR) models typically address the capture probabilities depending on the location of traps relative to animals. However, majority of the literature in this area uses the traditional SECR approach related to observing the traps state only at specific intervals of time, e.g. at the end of each day. Nowadays traps are usually substituted with cameras that can provide detailed temporal occurrence information. Hence, we assume that the data is observed in continuous time which allows us to use them in a typical point process manner. Self-exciting concept appears reasonable since animal movement has sequential inherent pattern. However, the data for each animal can be sparse which we overcome by hierarchical pooling, since all animals from a certain species share similar behavioural patterns. Hawkes process is usually applied in frequentist's manner which ignores the inherent uncertainty that arises as part of the model structure. Hence, future prediction based on a historical catalogue can result in misleading forecasts that heavily under or overestimate the process' multi-layered intensity structure. Bayesian statistics allows parameter uncertainty to be explicitly represented and used afterwards for obtaining a forecast distribution. Hence, we developed hybrid-multivariate spatial Hawkes process and we further addressed its estimation in a Bayesian manner based on the underlying hidden inheritance structure.

Spatial Analysis of Laser-Induced Breakdown Spectroscopy Data for a Sandstone Sample

Tuesday 23
16:30 - 17:30
Second Floor

Tereza Konečná[1]

[1]Institute of Mathematics FME BUT, Brno University of Technology, Brno, Czech Republic
151717@vutbr.cz

This paper aims at an analysis of data obtained by Laser-Induced Breakdown Spectroscopy (LIBS) of a sandstone sample. LIBS is a device for the implementation of qualitative and quantitative analyses by a laser pulse (with better spatial resolution and detection limits compared to other classical methods). A spectrum of received intensities at various wavelengths is obtained for each of measuring points in a dense regular grid with a constant distance between them and chemical composition of the sample in each point can be estimated from the spectrum. In case of modelling intensities of selected chemical elements in the neighbouring points, spatial models allow for dependence among the explained variables and for covariates in the sense of other chemical elements can be applied with an advantage. Here we aim at modelling a specific relation between the presence of uranium and other chemical elements as covariates in a studied sample of uraniumhosted sandstone. Obtained models can be used, for example, to decrease the density of the measurement grid and thereby accelerate the measurement process and reduce the sample contamination. Computational models comprise the methods of universal kriging and variogram model fitting. Software implementations of mentioned models are programmed using the packages geoR and geoRglm.

Statistical Analysis of the Impact of Air Pollution on the Circulatory System

Tuesday 23
16:30 - 17:30
Second Floor

Dominika Korbas[1] and Barbara Jasiulis-Gołdyn

[1]Mathematical Institute, Univeristy of Wrocław, Wrocław, Poland
dominique065@gmail.com

The World Health Organization considers pollen mixtures PM2.5 to be the most harmful to health among other types of atmospheric pollution. Dust particles penetrates directly into the lungs and circulatory system, from where toxic compounds are transferred into the whole organism. Our main goal is to analyze the impact of air pollution indicators and meteorological data on the circulatory system based on number of ambulance calls for cardiovascular diseases in Wrocław (Poland). Finally, using R programming, we are going to create a model to predict the number of beds in hospitals in Wrocław for patients with cardiovascular diseases.

Tuesday 23
14:00 - 16:00
Room 7

Bootstrap Based Inference for Sparse High-Dimensional Time Series Models

Jonas Krampe[1] and Jens-Peter Kreiss and Stathis Paparoditis

[1]Department of Economics, University of Mannheim, Mannheim, Germany
j.krampe@uni-mannheim.de

Fitting sparse models to high dimensional time series is an important area of statistical inference. In this paper we consider sparse vector autoregressive models and develop appropriate bootstrap methods to infer properties of such processes.

Our bootstrap methodology generates pseudo time series using a model-based bootstrap procedure which involves an estimated, sparsified version of the underlying vector autoregressive model. Inference is performed using so-called de-sparsified or de-biased estimators of the autoregressive model parameters. We derive the asymptotic distribution of such estimators in the time series context and establish asymptotic validity of the bootstrap procedure proposed for estimation and, appropriately modified, for testing purposes. In particular we focus on testing that groups of autoregressive coefficients equal zero. Our theoretical results are complemented by simulations which investigate the finite sample performance of the bootstrap methodology proposed. A real-life data application is also presented.

Thursday 25
14:00 - 16:00
Room 11

Functional Data Analysis and Censoring

David Kraus[1] and Stanislav Nagy

[1]Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic
david.kraus@mail.muni.cz

Functional data analysis is often complicated by the fact that the information collected about the curves is distorted due to incomplete, fragmentary, discrete, noisy or otherwise imperfect observation. In this talk we focus on censoring. We discuss the estimation of characteristics of the distribution of functional data and prediction of functional observations.

Advances on the Asymptotic Normality of Betti Numbers

Tuesday 23
10:30 - 12:30
Room 12

Johannes Krebs[1] and Wolfgang Polonik

[1]Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany
jtkrebs@ucdavis.edu

We study persistent Betti numbers in the critical regime. Persistent Betti numbers are a major tool in persistent homology (topological data analysis), which is a multiscale approach to quantifying topological features in data, in particular, point cloud data. First, we present the basic ideas of topological data analysis and persistent homology. We introduce persistent Betti numbers and the corresponding persistence diagram. Then we consider persistent Betti numbers from a Poisson and a binomial sampling scheme. We study their strong stabilizing property and their pointwise asymptotic normality in the critical regime. Also we discuss the generalization towards the weak convergence of two parameter process which describes the persistent Betti function. Moreover, we address potential applications in data analysis such as the bootstrap and change point detection.

Statistical Inference on Networks: Correlation Bounds, Mixing and M-Dependence under Random, Time-Varying Network Distances with Applications to a Dynamic Network Model Based on Counting Processes

Monday 22
14:00 - 16:00
Room 8

Alexander Kreiß[1]

[1]Deparment of Economics, University of Mannheim, Mannheim, Germany
kreiss@uni-mannheim.de

In the present work we will consider multivariate stochastic processes indexed either by vertices or edges of a dynamic network. The dynamics of the network is allowed to be random. We will assume that the spatial dependence-structure of the processes is linked with the network in the following way: Two neighbouring vertices (or two adjacent edges) are dependent, while we assume that the dependence decreases as the distance in the network increases. We make this intuition mathematically precise by considering three concepts based on correlation, β-mixing (with time-varying β-coefficients) and conditional independence. Then, we will use these concepts in order to prove weak-dependency results (e.g. an exponential inequality) which might be of independent interest. In order to demonstrate the use of these concepts in an application we study a goodness of fit test in a dynamic interaction network model based on a multiplicative hazard model. This model is then applied to bike-sharing data.

Monday 22
14:00 - 16:00
Room 10

Adaptive Non-Parametric Estimation of Mean and Autocovariance in Regression with Dependent Errors

Tatyana Krivobokova[1] and Paulo Serra

[1]University of Göttingen, Göttingen, Germany
tkrivob@gwdg.de

We develop a fully automatic non-parametric approach to simultaneous estimation of mean and autocovariance functions in regression with dependent errors. Our empirical Bayesian approach is adaptive, numerically efficient and allows for the construction of confidence sets for the regression function. Consistency of the estimators is shown and small sample performance is demonstrated in simulations and real data analysis.

Monday 22
10:30 - 12:30
Room 12

Local Differential Privacy: Elbow Effect in Optimal Density Estimation and Adaptation over Besov Ellipsoids

Martin Kroll[1] and Cristina Butucea and Amandine Dubois and Adrien Saumard

[1]CREST, ENSAE, Institut Polytechnique de Paris, Palaiseau, France
martin.kroll@ensae.fr

We address the problem of non-parametric density estimation under the additional constraint that only privatised data are allowed to be published and available for inference. For this purpose, we adopt a recent generalisation of classical minimax theory to the framework of local α-differential privacy and provide a lower bound on the rate of convergence over Besov spaces B^s_{pq} under mean integrated $\mathbb{L}^r$-risk. This lower bound is deteriorated compared to the standard setup without privacy, and reveals a twofold elbow effect. In order to fulfil the privacy requirement, we suggest adding suitably scaled Laplace noise to empirical wavelet coefficients. We derive upper bounds within (at most) a logarithmic factor of the upper bounds under the assumption that α stays bounded as n increases: A linear but non-adaptive wavelet estimator is shown to attain the lower bound whenever $p \geq r$ but provides a slower rate of convergence otherwise. An adaptive non-linear wavelet estimator with appropriately chosen smoothing parameters and thresholding is shown to attain the lower bound within a logarithmic factor for all cases.

Wednesday 24
10:30 - 12:30
Seminari B

Estimation of the Fisher Information Matrix in Latent Variables Models Based on the Score Function

Estelle Kuhn[1] and Maud Delattre

[1]Applied Mathematic, French National Institute for Agricultural Research, Jouy-en-Josas, France
estelle.kuhn@inra.fr

The Fisher information matrix (FIM) is a key quantity in statistics as it is required for examples for evaluating asymptotic precisions of parameter estimates, for computing asymptotic distributions in statistical testing or in post model selection inference. However its exact computation is often not trivial. In particular in many latent variable models, it is intricated due to the presence of unobserved variables. Therefore the observed FIM is usually considered. Several methods have been proposed to approximate the observed FIM when it cannot be evaluated analytically. Let us quote for examples Monte-Carlo methods or iterative algorithms derived from the missing information principle. All these methods require to compute second derivatives of the complete log-likelihood leading to possible computational limit. In this talk, we present a new approach to estimate the FIM in latent variable model. Its advantage is that only the first derivatives of the complete log-likelihood is needed. Indeed we consider the empirical estimate of the covariance matrix of the score. We prove that this estimate of the Fisher information matrix is unbiased, consistent and asymptotically Gaussian. When the proposed estimate cannot be exactly evaluated, we present a stochastic approximation estimation algorithm to compute it. This algorithm provides this estimate of the FIM as a by-product of parameter estimates. We prove that the estimation algorithm is consistent and asymptotically Gaussian when the number of iterations goes to infinity. We evaluate the finite sample size properties of both estimates through simulation studies in linear mixed effects models and mixture models. We also investigate the convergence properties of the estimation algorithm in nonlinear mixed effects models and frailty models. We compare the performances of the proposed algorithm to those of other existing methods. Finally we apply the procedure to analyse mastitis infection data using frailty models.

What is the Resolution of a Microscope? A Statistical Minimax Point of View

Thursday 25
14:00 - 16:00
Room 6

Gytis Kulaitis[1] and Axel Munk and Frank Werner

[1]Institute for Mathematical Stochastics, Goettingen University, Goettingen, Germany
gytis.kulaitis@mathematik.uni-goettingen.de

A general rule of thumb in imaging is that the resolution of a light microscope (i.e. the ability to discern objects) is described by the full width at half maximum (FWHM) of its point spread function (PSF) and depends linearly on it. This is based on physical reasoning founded in wave optics. However, with increasing resolution the random nature of observations (light photons) becomes increasingly important. This calls for a statistical description of resolution obtained from such random data which we provide in this paper for the first time. Our approach is based on testing whether one or two objects with the same total intensity are present. For Poisson or binomial random variables we get linear dependence on the FWHM, however for Gaussian random variables the dependence of resolution is to the power of 5/4. This shows that modeling resolution with Gaussian random variables is inadequate, whereas Poisson and binomial are sufficient to describe resolution mathematically.

Monday 22
16:30 - 18:30
Room 5

Some results on the linear combination of chi-squares with applications to inference in shape and directional statistics

Alfred Kume[1] and Tomonari Sei and Andrew A. T. A. Wood

[1]School of Mathematics, Statisticias and Actuarial Science, University of Kent, Canterbury, United Kingdom
a.kume@kent.ac.uk

Some random matrix models adopted for statistical analysis of directions and shapes, rely on certain relationships with the linear combination of central and non-central chi-square random variables. Motivated initially from the directional statistics problems, we focus on the density function of such distributions and not on their cumulative distribution function which have been extensively covered in the literature.

Our approach provides new insight by generating alternative characterisations of the relevant expressions for a range of distributions used in directional and shape inference. In addition, our results can be easily extended to some general expectations used for spike models used in random matrix theory. The expressions we obtain are more transparent for modelling purposes and with apparent stability in their numerical evaluation. We will illustrate our method with some examples.

Tuesday 23
10:30 - 12:30
Room 6

Efficient Computation of Pseudo-Gibbs Distributions

Kun-Lin Kuo[1]

[1]Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan
klkuo@nuk.edu.tw

The pseudo-Gibbs sampling (PGS) method is often used to generate samples and use the empirical distribution to approximate joint distributions, though its properties are mostly unknown. Kuo and Wang (2019) investigate the richness and varieties of the joint distributions that PGS approximates. In short, PGS produces a large number of joint distributions with shared dependence and shared marginal distributions. As an algorithm, PGS has multiple fixed points that are as similar as possible under the constraints of incompatibility; the optimal similarity is proved in terms of Kullback-Leibler information divergence. Moreover, properties of incompatible PGS provide a fresh perspective for understanding the original Gibbs sampler used in Bayesian computations. In this talk, we provide an efficient computation for pseudo-Gibbs distributions.

Using Randomization Testing to Estimate the Effects of Gerrymandering in US Elections

Tuesday 23
14:00 - 16:00
Room 8

Paul Kvam[1] and Daniel Palazzolo

[1]Mathematics and Computer Science, University of Richmond, Richmond, United States
pkvam@richmond.edu

Political scientists find that the effects of partisan gerrymandering on election outcomes for the US House of Representatives can be confounded with the geographical distribution of Democratic and Republican voters. We use simple ideas from statistical randomization tests to show how the effects of these geographic distributions pertain to election outcomes. As a result, we are more able to differentiate those effects from the less ambiguous outcomes of gerrymandering in congressional elections. Using data from the 2016 presidential campaign, we find varied effects of partisan redistricting at the state level among states where the legislatures allegedly engaged in partisan gerrymandering. In any case, the natural distribution of voters prevents Democrats from translating a majority of the popular vote into a majority of House seats. In simulated redistricting outcomes nationwide, Democrats rarely achieve more than 45

NAR-Sieve Bootstrap Based on Extreme Learning Machines

Tuesday 23
14:00 - 16:00
Room 7

Michele La Rocca[1] and Cira Perna

[1]Dept. Economics and Statistics, Salerno, Italy
larocca@unisa.it

In the last decade, Extreme learning machines (ELM) has become a very popular learning tool for Single-hidden Layer Feedforward Networks. In this algorithm, the input weights and the hidden layer bias are randomly chosen with remarkable computational advantages with respect to standard neural network models. The ELM formulation leads to solving a system of linear equations in terms of the unknown weights connecting the hidden layer to the output layer. The solution of this general system of linear equations is usually obtained using Moore-Penrose generalised pseudoinverse. The aim of the talk is to propose and discuss a sieve bootstrap scheme based on ELM for nonlinear time series. The procedure is fully nonparametric in its spirit and retains the conceptual simplicity of the AR-Sieve bootstrap. Using ELMs in the resampling scheme can dramatically reduce the computational burden of the bootstrap procedure, with performances comparable to the NN-Sieve bootstrap and computing time similar to the AR-Sieve bootstrap. Moreover, NAR-Sieve bootstrap could be easily extended to multivariate nonlinear time series making this bootstrap scheme suitable for the construction of inference tools for complex vector time series.

Friday 26
11:30 - 12:30
Aula Magna
Building 13

Computational Perspectives on some Statistical Problems

John Lafferty[1]

[1]Department of Statistics and Data Science, Yale University, USA
john.lafferty@yale.edu

We present some variations on classical statistical problems that take a computational, machine learning perspective. First, we study nonparametric regression when the data are distributed across multiple machines. We place limits on the number of bits that each machine can use to transmit information to the central machine. Our results give both asymptotic lower bounds and matching upper bounds on the statistical risk under various settings. Second, we investigate the use of machine learning algorithms that are required to obey natural shape constraints suggested by domain knowledge. We develop methods for high-dimensional shape-constrained regression and classification that "reshape" the original prediction rule. Finally, we study optimization procedures to minimize the empirical risk functional for certain families of deep neural networks. We develop an approach that optimizes a sequence of objective functions using network parameters obtained during different stages of the learning process. This is evaluated with deep generative networks used as a replacement for sparsity in compressed sensing and approximation.

Friday 26
09:00 - 11:00
Room 8

Likelihood-Free Inference of Multi-Scale Mathematical Model for Personalised Breast Cancer Treatment

Xiaoran Lai[1] and Henri Pesonen and Jukka Corander and Arnoldo Frigessi

[1]Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway
xiaoran.lai@medisin.uio.no

Multi-type routinely acquired measurements are usually noisy and are available at only aggregated level. Given the quality and granularity of these longitudinal data, one of the major difficulties in using mechanistic models to accurately predict individual treatment outcomes lies in the uncertainty of model parameters. Using Approximate Bayesian Computation (ABC) allows us to improve accuracy of the estimation based on available data and prior knowledge. In this study, we investigated a complex multi-scale model of breast cancer treated with chemotherapeutic and anti-angiogenic agents. The model contains biologically-related parameters such as cell proliferation and vessel perfusion. Our results demonstrated that ABC can be used i) to estimate previously calibrated parameters given prior information, ii) to provides non-linear alternative therapies with uncertainty, and iii) to guide better experiment design. We also identified ways to improve efficiency of the estimation and discussed extension of joint estimation of dependent parameters.

A Missing Value Approach for Breeding Value Estimation

Thursday 25
10:30 - 12:30
Room 8

Alastair Lamont[1] and Matthew Schofield and Richard Barker

[1]University of Otago, Dunedin, New Zealand
alamont@maths.otago.ac.nz

Breeding value is a measure of how valuable an individual's genetic material is when inherited by its progeny. Accurate breeding value estimation is a critical component of selective breeding, allowing prediction of future offspring. Genetic data is commonly available and can be used to improve breeding value estimation. This genetic data is often only partially observed, with some animals unmeasured. Modern approaches adapt least-squares estimation to allow for unobserved genetic data. To do so requires particular assumptions and approximations which may not be well-suited to typical livestock data. Breeding values can also be estimated using Bayesian methods, but existing approaches do not accommodate missing genetic data. We specify an Bayesian approach which estimates missing genotypes alongside other parameters. This is an extension of several similar approaches used in human applications. An approximation to this approach has the potential to greatly improve computational speed with little cost in accuracy.

Generalizing Stochastic Orders Through Probability Transformation Functions

Friday 26
09:00 - 11:00
Seminari B

Tommaso Lando[1] and Lucio Bertoli-Barsotti

[1]University of Bergamo and VSB Technical University of Ostrava, Bergamo, Italy
tommaso.lando@unibg.it

We study a semi-parametric family of stochastic orders, modeled by a transformation of the cumulative probability distribution, generally referred to as probability distortion. Such class determines a continuum of dominance relations, where the classic first and second-degree stochastic dominance may be obtained as a limiting case and a special case, respectively. In the theory of decision under uncertainty, this general approach makes it possible to represent the preferences of decision makers who may exhibit some degree of flexibility in terms of risk attitude/risk aversion.

Tuesday 23
14:00 - 16:00
Room 12

Aggregated Tests of Independence Based on HSIC Measures

Béatrice Laurent[1] and Anouar Meynaoui and Mélisande Albert and Amandine Marrel

[1]Institut de Mathématiques de Toulouse, INSA de Toulouse, Toulouse, France
beatrice.laurent@insa-toulouse.fr

Dependence measures based on reproducing kernel Hilbert spaces, also known as Hilbert-Schmidt Independence Criterion and denoted HSIC, are widely used to statistically decide whether or not two random vectors are dependent. Recently, non-parametric HSIC-based statistical tests of independence have been performed. However, these tests lead to the question of prior choice of the kernels associated to HSIC, there is as yet no method to objectively select specific kernels. In order to avoid a particular kernel choice, we propose a new HSIC-based aggregated procedure allowing to take into account several Gaussian kernels. To achieve this, we first propose non-asymptotic single tests of level $\alpha \in (0, 1)$ and second type error controlled by $\beta \in (0, 1)$. We also provide a sharp upper bound of the uniform seperation rate of the proposed tests. Thereafter, we introduce a multiple testing procedure in the case of Gaussian kernels, considering a set of various parameters. These agregated tests are shown to be of level α and to overperform single tests in terms of uniform separation rates.

We propose two contributed talks, one after the other. The first talk from Béatrice Laurent will present a single test based on HSIC measures and its non asymptotic performances. The second talk given by Anouar Meynaoui will present the aggregated procedures, their theoretical performances and the simulation study.

Tuesday 23
10:30 - 12:30
Seminari B

Bayesian Stochastic Search for Ising Chain Graph Models

Andrea Lazzerini[1] and Monia Lupparelli and Francesco C. Stingo

[1]University of Bologna, Bologna, Italy
andrea.lazzerini2@unibo.it

A class of Ising chain graph models is illustrated to explore the effect of an external risk factor on a set of binary outcomes and on their joint dependence structure modelled via undirected graphs, where we are mainly interested in the risk factor effect on pairwise associations between outcomes rather on single outcomes. Under the LWF Markov property, the joint probability mass associated to a chain graph corresponds to a log-linear model with suitable zero constraints in correspondence of missing edges. We devise a Bayesian Ising model based on conjugate priors for log-linear parameters that aims at the selection of the best graph that fits the data. A computational strategy is implemented that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform model selection

---

Limiting Behaviour of Frechet Means on Manifolds

Monday 22
16:30 - 18:30
Room 5

Huiling Le[1]

[1]School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom
huiling.le@nottingham.ac.uk

The concept of Frechet means, as a generalisation of Euclidean means, is becoming an increasingly important fundamental tool in many statistical analyses of data of a non-Euclidean nature. However, the non-standard probabilistic and statistical features of Frechet means give rise to many challenges, both practical and mathematical. A good understanding of these features is crucial for developing appropriate statistical methods required, for example, for computation, estimation and inference. In this talk, I will discuss some recent progress in understanding the interplay between the probabilistic behaviour of Frechet means and the geometric structure of Riemannian manifolds.

Fast Bayesian Inference in Large Graphical Models

Tuesday 23
10:30 - 12:30
Room 9

Gwenaël G.R. Leday[1] and Sylvia Richardson

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom
gwenael.leday@mrc-bsu.cam.ac.uk

Despite major methodological developments, Bayesian inference in Gaussian graphical models remains challenging in high dimension due to the tremendous size of the model space. In this talk, we propose a method to infer the conditional independence structure between variables by multiple testing, which bypasses the exploration of the model space. Specifically, we introduce closed-form Bayes factors under the Gaussian conjugate model to evaluate the null hypotheses of marginal and conditional independence between variables. Their computation for all pairs of variables is shown to be extremely efficient, thereby allowing us to address large problems with thousands of nodes as required by modern applications. Moreover, we derive exact tail probabilities from the null distributions of the Bayes factors. These allow the use of any multiplicity correction procedure to control error rates for incorrect edge inclusion. We demonstrate the proposed approach on various simulated examples as well as on a large gene expression data set from The Cancer Genome Atlas.

Testing for Change-Point in the Covariate Effects Based on the Cox Proportional Hazards Model

Monday 22
14:00 - 16:00
Seminari B

Chun Yin Lee[1] and Xuerong Chen and K. F. Lam

[1]Department of Statistics and Actuarial Sciences, The University of Hong Kong, Hong Kong
jacylee@hku.hk

Models with change-point in covariates have wide applications in cancer research with the response being the time to a certain event. A Cox model with change-point in covariate is considered at which the pattern of the change-point effects can be flexibly specified. To test for the existence of the change-point effects, three statistical tests, namely the maximal score, maximal normalized score and maximal Wald tests are proposed. The asymptotic properties of the test statistics are established. A Monte Carlo approach to simulate the critical values is suggested. A large-scale simulation study is carried out to study the performance of the proposed test statistics under the null hypothesis of no change-point and various alternative hypothesis settings. Each of the proposed methods provide a natural estimate for the location of the change-point, but it is found that the performance of the maximal score test can be sensitive to the true location of the change-point in some cases, while the performance of the maximal Wald test is very satisfactory in general even in cases with moderate sample size. For illustration, the proposed methods are applied to three medical datasets concerning patients with primary biliary cirrhosis, glioblastoma multiforme and breast cancer, respectively.

Tuesday 23
16:30 - 17:30
Second Floor

Approximate Maximum Product of Spacing Estimation of the Parameter for a Half-Logistic Distribution Based on Progressive Censoring

Kyeongjun Lee[1] and Subin Cho and Yeongeun Hwang

[1]Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, Korea, South
indra_74@naver.com

Recently, progressive censoring has become quite popular in a life-testing problem and reliability analysis. Under classical estimation set up, the maximum product spacings method is quite effective and several authors advocated the use of this method as an alternative to maximum likelihood estimation method, and found that this estimation method provides better estimates than maximum likelihood estimate in various situations. In this paper, we derive the maximum product spacings estimator for the parameter and reliability function of half-logistic distribution. And we derive the approximate maximum product spacings estimator for the parameter and reliability function of half-logistic distribution using Talyor series expansion. We also compare the proposed estimators in the sense of the root mean squared error RMSE and bias for various progressive censoring schemes. In addition, real data example based on progressive censoring have been also analysed for illustrative purposes.

Local White Noise Testing with Wavelets

Friday 26
09:00 - 11:00
Room 12

Kathryn Leeming[1] and Guy Nason

[1]Statistical Science, University of Bristol, Bristol, United Kingdom
kathryn.leeming@bristol.ac.uk

We present a new white noise test designed to be applied locally to time series. Most white noise tests are designed for global use, where the entire series is tested at once, but our test can be applied locally at multiple points of a time series. Such time series may be observations, or residuals from model fitting. In this talk we motivate the use of wavelets for local white noise testing, and present theoretical results about the distribution of the wavelet spectrum.

Alongside spectrum results, we also discuss a choice of test statistics based upon quadratic forms and show the effectiveness of our local white noise test variants against alternative hypotheses. In addition, examples show the benefit of using a local white noise test in situations where global tests falsely fail to reject the null hypothesis.

Functional Data Clustering and Missing Value Imputation

Tuesday 23
16:30 - 17:30
Second Floor

Pai-Ling Li[1]

[1]Department of Statistics, Tamkang, New Taipei City, Taiwan
plli@gms.tku.edu.tw

We propose a functional data approach to clustering and missing value imputation for incomplete longitudinal data. We adopt the notion of subspace-projected functional data clustering that each observed trajectory is viewed as a realization of a random function and is drawn from a mixture of stochastic processes, where each subprocess represents a cluster with a cluster-specific mean function and covariance function. The proposed algorithm comprises the probabilistic functional clustering and the missing value imputation based on the clustering results. The performance of the proposed method is demonstrated through a data example.

A Precise Framework for Testing Weak Instruments in High-Dimensional Instrumental Variables Regression

Monday 22
10:30 - 12:30
Room 8

Zhaoyuan Li[1] and Jianfeng Yao

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China
lizhaoyuan@cuhk.edu.cn

This paper considers tests of weak instruments in high-dimensional instrumental variable regression where the numbers of endogenous variables and instruments can be large compared to sample size. The focus is the first-stage non-Fisher matrix. The limiting spectral distribution and central limit theory of its eigenvalues are established based on random matrix theory. Then, a precise framework for weak instruments testing is constructed. An empirical example and Monte Carlo evidence demonstrate the superiority of the proposed tests over popular existing methods. As expected, the proposed tests perform very well under both the high-dimensional setting and the classical low-dimensional situation.

Tuesday 23
10:30 - 12:30
Room 10

A New Multivariate Analysis Framework to Investigate Complex Interaction Patterns

Ker-Chau Li[1]

[1]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
kcli@stat.sinica.edu.tw

We present a new statistical framework to investigate complex interaction between variables. Our study is motivated by the lack of a systematic approach for modeling data generated by high throughput biomolecular assays. While most computational tools for exploring voluminous biological data are based on pair-wise correlation between variables, we have earlier introduced Liquid Association (LA) for inferring higher order of association between variables in a system (Li 2002, Proceedings of National Academy of Sciences, USA). LA was originally introduced to study patterns of gene-gene interaction that involve three genes at a time. LA aims at finding how the correlation pattern between a pair of functionally associated genes may be dynamically altered due to the influence of ever-changing but often-hidden cellular state. It is computationally expensive to compute LA scores for all possible triplets in very large datasets, but Graphic Processing Units (GPUs) can be used to speed up the LA computing. In this talk, we shall show how LA, in junction with several ways of extension such as background adjustment and dimension reduction, can be used to broaden the scope of multivariate analysis.

Thursday 25
14:00 - 16:00
Room 7

Quantile-Frequency Analysis and Functional Principal Components for Time Series Classification

Ta-Hsin Li[1]

[1]IBM Research, Yorktown Heights, United States
thl@us.ibm.com

Quantile periodogram is a recently proposed tool for spectral analysis of time-series data. Derived from trigonometric quantile regression, the quantile periodogram offers a capability of quantile-frequency analysis (QFA) that characterizes the oscillatory behavior of time series around different quantile levels. This talk introduces a QFA-based functional principal component analysis (FPCA) method that extracts useful features for discriminant analysis of time series. A real-world dataset of nondestructive evaluation (NDE) of

mechanical systems is used to demonstrate the advantages of the proposed method over the traditional spectral analysis method based on the ordinary periodogram.

Estimating APGARCH-Skew-T Model by Wavelet Support Vector Machines

Tuesday 23
14:00 - 16:00
Room 8

Yushu Li[1]

[1]Dep. of Mathematics, University of Bergen, Bergen, Norway
Yushu.li@uib.no

This paper concentrates on comparing estimation and forecasting ability of quasi-maximum likelihood (QML) and support vector machines (SVM) for financial data. The financial series are fitted into a family of asymmetric power ARCH (APARCH) models. As the skewness and kurtosis are common characteristics of the financial series, a skew-t distributed innovation is assumed to model the fat tail and asymmetry. Prior research indicates that the QML estimator for the APARCH model is inefficient when the data distribution shows departure from normality, so the current paper utilizes the semi-parametric-based SVM method and shows that it is more efficient than the QML under the skewed Student's-t distributed error. As the SVM is a kernel-based technique, we further investigate its performance by applying separately a Gaussian kernel and a wavelet kernel. The results suggest that the SVM-based method generally performs better than QML for both in-sample and out-of-sample data. The outcomes also highlight the fact that the wavelet kernel outperforms the Gaussian kernel with lower forecasting error, better generation capability and more computation efficiency

---

Design and Implementation of Recommender Based on Debiased Learning to Rank

Thursday 25
14:00 - 16:00
Room 6

Bo Li[1]

[1]Management Science and Engineering, Tsinghua University, Beijing, China
libo@sem.tsinghua.edu.cn

In this work, we adapt modern machine learning techniques to a classical decision problem with extra complications. Data-diven Markovitz portfolio optimization models face certain difficulties like estimation error and transaction costs when they are applied in practice. We consider mean-variance optimization problem accounting for transaction costs. To tackle the estimation errors, we employ two types of strategies, the performance-based regularization and the distributionally robust optimization. Comparative experiments based on real data sets are conducted and the experiment results demonstrate the competitive performance of the proposed methods.

Thursday 25
14:00 - 16:00
Room 11

Reconstructing Partially Observed Functional Data with (Non-)Systematically Missing Parts

Dominik Liebl[1] and Alois Kneip and Stefan Rameseder

[1]University Bonn, Bonn, Germany
dliebl@uni-bonn.de

The first part of the talk considers the case of partially observed functional data with non-systematically missing parts. A new reconstruction operator is proposed which aims to recover the missing parts of a function given the observed parts. This new operator belongs to a new, very large class of functional operators which includes the classical regression operators as a special case. The optimality of our reconstruction operator is shown and it is demonstrated that the usually considered regression operators generally cannot be optimal reconstruction operators. The estimation theory allows for autocorrelated functional data and considers the practically relevant situation in which the functions are only at noise-contaminated discretization points. Rates of consistency are derived for the nonparametric estimation procedures using a double asymptotic. The second part of the talk proposes new estimators for the mean and the covariance function for partially observed functional data using a detour via the fundamental theorem of calculus. These new estimators allow for a consistent estimation of the mean and covariance function under specific violations of the missing-completely-at-random assumption.

Friday 26
09:00 - 11:00
Seminari B

Extremality Order and Extremes Detection in the Cantabrian Coast

Rosa Lillo[1] and Raul Torres

[1]UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain
rosaelvira.lillo@uc3m.es

The extremality stochastic order and the idea of directional quantile allow the detection of multivariate extremes for which the direction in which it is more convenient (depending on the problem) to look at the data plays a fundamental role. In this paper we present a methodology to detect extremes that is scalable for high dimensions and that offers more credible results than those provided by the usual directions given by the survival function or distribution function. This methodology is illustrated with data that come from the Cantabrian coast of Spain.

Empirical Maximization of R-Statistics in the Two-Sample Problem and Nonparametric Homogeneity Tests in High Dimension

Monday 22
10:30 - 12:30
Room 8

Myrto Limnios[1] and Stephan Clémençon and Marine Depecker and Nicolas Vayatis

[1]CMLA, ENS Paris-Saclay, Paris, France
myrto.limnios@gmail.com

The ROC curve is the gold standard for measuring performance of a test/scoring statistic regarding its capacity to discriminate between two statistical populations in a wide variety of applications as anomaly detection in signal processing, information retrieval, medical diagnosis, etc. Most practical performance measures used in scoring/ranking applications such as the AUC, the local AUC, the p-norm push, the DCG and others, can be viewed as summaries of the ROC curve. It is the first goal of this paper to highlight the fact that most of these empirical criteria can be expressed as linear rank statistics and next investigate the properties of empirical maximizers of such performance criteria in a general framework. Preliminary results for the concentration properties of a novel class of random variables that we call linear rank processes are established for this purpose. Secondly, the paper explores the connection between multivariate homogeneity tests and optimization of the performance criteria aforementioned. From the elementary observation that, in the two-sample problem, the null assumption corresponds to the situation where the optimal ROC curve coincides with the first diagonal of the unit square, we propose a two-stage testing method based on data splitting. A nearly optimal scoring function in the sense of a linear rank criteria is first learnt from one of the two half-samples. Data from the remaining half-sample are then projected onto the real line and eventually ranked according to the scoring function computed at the first stage. The last step performs a standard rank test in the one-dimensional framework. We show that the learning step of the procedure does not affect the consistency of the test as well as its properties in terms of power, provided the ranking produced is accurate enough. The results of numerical experiments are also displayed in order to illustrate the efficiency of the method, compared to that of alternative testing procedures.

Modeling Financial Interval Time Series

Tuesday 23
10:30 - 12:30
Room 5

Liang-Ching Lin[1] and Li-Hsien Sun

[1]National Cheng Kung University, Tainan, Taiwan
lclin@mail.ncku.edu.tw

In financial economics, a large number of models are developed based on the daily closing price. When using only the daily closing price to model the time series, we may discard valuable intra-daily information, such as maximum and minimum prices. In this study, we propose an interval time series model, including the daily maximum, minimum, and closing prices, and then apply the proposed model to forecast the entire interval. The likelihood function and the corresponding maximum likelihood estimates (MLEs) are obtained by stochastic differential equation and the Girsanov theorem. To capture the heteroscedasticity of volatility, we consider a stochastic volatility model. The efficiency of the proposed estimators is illustrated by a simulation study. Finally, based on real data for S&P 500

index, the proposed method outperforms several alternatives in terms of the accurate forecast.

Tuesday 23
16:30 - 17:30
Second Floor

Modified Information Matrix Tests for Detecting Misspecification in the Random-Effects Structure of Cumulative Logit Models

Kuo-Chin Lin[1]

[1]Department of Business Administration, Tainan University of Technology, Tainan
t20053@mail.tut.edu.tw

A common approach for analyzing longitudinal ordinal data is to apply generalized linear mixed models (GLMMs). The normality assumption of the random-effects distribution in cumulative logit mixed models is practically assumed, but it may be too restrictive to reveal the major feature of data. The test statistics for normality assumption are proposed based on a variety of modified information matrix tests introduced by White (1982), and their limiting chi-squared distributions are derived under the null hypothesis that the distribution of random effects is corrected specified. Simulation results are presented under various configurations of practical relevance data generating mechanism with different modified matrices and sample sizes, and the power performance of the proposed tests is demonstrated. Furthermore, a respiratory disorder study with longitudinal ordinal responses is employed to illustrate the applications of proposed tests.

Friday 26
09:00 - 11:00
Room 7

Build Imbalanced Classifier via Kernel Trick and Evaluate Its Effect

Wei Hsiang Lin[1] and Charlotte Wang

[1]Mathematics, Tamkang University, New Taipei City, Taiwan
kusoboy5202@gmail.com

In the age of big data,class imbalance data is a common problem in actual application.When the data with skewed distributions,the sample size of one class is much more than other class .So the traditional methods are more focusing on classification of major sample while ignoring or misclassifying minority sample.It's a challenge to find a suitable method.In last few year there many methods proposed that kernel transformation is useful for classification.In this paper,we want to explore what kind of kernel transformation corresponding to different datasets that can maintain geometric feature of original data and then we establish an suitable classification model by ensemble method.Finally, we propose a method to sort out imbalance problem and verify by real data.

The Determinants of Food Waste: Evidence from a Semiparametric Functional-Coefficient Cointegration Test

Tuesday 23
16:30 - 17:30
Second Floor

Yi-Chen Lin[1] and Wen-shuenn Deng and Kaizhi Yu and Xiang Chen

[1]Department of Economics, Tamkang University, Tamsui, Taiwan
ylin@mail.tku.edu.tw

This paper applies a conventional linear cointegration test and a more flexible semiparametric varying-coefficient cointegration test to a unique time series of household food waste at the city level over the period 2003m1 to 2018m8. Both tests support the presence of a level relationship among the extent of waste, as measured by the edible-to-inedible food scrap ratio, food price, population age structure, average household size, and average real labor income. We find that a 1% increase in food price, share of population aged 15 to 65 years, share of population aged 65 or over, and household size results in a 1.227%, -2.358%, 0.723%, and 5.684% change in the long-run level of the edible-to-inedible food scrap ratio, respectively. The semiparametric varying-coefficient model further reveals that the direct relationship between average income and the extent of waste is an inverted U. The elasticity of the edible-to-inedible food scrap ratio with respect to food price and the share of working (old)-age population are negative (positive) at high income levels. Our empirical results imply that food waste is likely to grow continuously with an ageing population, but it is also likely to fall with a long-term reduction in young people's household formation

---

Covariance Estimation of Matrix-Variates Through Iterative Process

Monday 22
10:30 - 12:30
Room 8

Yiming Liu[1] and Pan Guangming, Leng Chenlei

[1]SPMS/Mathematics, Nanyang Technological Univeristy, Singapore
liuy0135@e.ntu.edu.sg

We introduce an iterative method to estimate the vectorized covariance matrix for matrices data and prove that such iterative process converges in a fast rate. Particularly, the iterative process always ends within finite iterations in practice. Our proposed method is applicable to a wider scope of different structures of the vectorized covariance matrix. Moreover, for the traditional assumption towards such covariance matrices, i.e., the Kronecker structure without any additional terms, the iterative method can be also applied. From our simulation studies, the performance of our proposed method also achieves to that under the Kronecker structure setting. Thus, our method is more robust and it is worthwhile to investigate. In addition, due to the special structure of the covariance matrix, our proposed "incoherence" condition is less restrictive in identifying two unknown matrices. We also show that we can estimate these two matrices separately. The performance of our estimators with other methods are conducted through extensive simulations and real data analysis.

Tuesday 23
10:30 - 12:30
Room 5

Recurrent Neural Network for Short-Term Load Forecasting with Spline Bases

Mong-Na Lo Huang[1] and Tzu-Lun Yuan and Chan-Nan Lu and Chin-Chung Wu

[1]Department of Applied Mathematics, National Sun Yat-Sen University, Kaohsiung, Taiwan
lomn@math.nsysu.edu.tw

In this work, we investigate how to make short term electricity load forecasting with the methodology of the recurrent neural network (RNN) combining the semi-parametric regression modeling. Performances of the RNN in short term load forecasting (STLF) are evaluated based on data obtained from the Taiwan Power Company (TPC) from January 2015 to December 2018. We build the RNN forecasting model with input variables from regressors of the semi-parametric regression model, such as the periodic B-spline bases related to the intra-daily and the intra-weekly effects, the historical actual temperature, as well as the forecasting temperature of the present day and seven days ahead forecasts provided by the Central Weather Bureau of Taiwan. The mean absolute percentage error (MAPE) is used for evaluation of the accuracies of the forecasting results. The performances of the RNN are presented and compared with some other methods such as a modified approach following artificial neural network (ANN) forecasting, and a semi-parametric regression model only approach. It is noted that training sample selection is very important for making accurate predictions of the load values. Clustering of the electricity loads based on the daily load patterns has also be investigated for the improvement of the forecasting especially during the weather changing periods.

Tuesday 23
16:30 - 17:30
Second Floor

Single Diusing Particles Observed Through a Confocal Mi-Croscope: An Application of the Doubly Stochastic Poisson Point Process.

Marco Longfils[1] and Magnus Röding and Mats Rudemo and Aila Särkkä

[1]Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden
longfils@chalmers.se

Diffusing particles observed with a confocal laser scanning microscope give rise to a doubly stochastic Poisson point process. In particular, the photon detected by the microscope in one pixel follows a Poisson distribution with parameter that depends on the particle positions in space, which is modelled as a Poisson point process. Many techniques such as Fluorescence correlation spectroscopy, Raster image correlation spectroscopy and photon counting histograms have been developed to study molecular transport in cells and solution. All these techniques are based on the statistics of the photon detection process. We show that the moments of the photon detection process can be computed in terms of physically relevant parameters such as the diffusion coefficient of the particles, their brightness and others. As a direct consequence, the statistical accuracy of the above mentioned techniques can be evaluated. Thus, we can relate the different experimental parameters that affects the photon detection process to the accuracy of each techniques, allowing us to optimally design an experiment.

Bayesian Inference for Parameter Estimation in High Dimension Space in Astrophysics

Thursday 25
14:00 - 16:00
Room 5

Javier Lopez-Santiago[1] and Luca Martino and Joaquín Miguez

[1]Dep. of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganes, Spain
jalopezs@ing.uc3m.es

Bayesian inference tools have become very popular in the Astrophysical community in recent years. One of the most common problems in Astronomy is to determine the value of the parameters of a model when only observations of several observables related to the model are available. In the past, least-squares or maximum likelihood methods were preferred. Nowadays, Monte Carlo methods are being used instead. In particular, MCMC algorithms are widely used for high dimensional problems. Nevertheless, some models require a particular treatment, particularly when the model has an undefined number of parameters a priori. Researchers are reporting difficulties in deriving a marginal likelihood that can be used to decide the actual number of parameters of the model. The problem of determining the number of exoplanets orbiting a star is a good example. From simulations of different stellar systems containing different number of planets, we have observed that a simple Bayesian information criterium may be used to discriminate between models.

Is-ClusterMPP: Clustering Algorithm Through Point Processes and Influence Space Towards High-Dimensional Data

Wednesday 24
10:30 - 12:30
Room 8

Pierre-Yves Louis[1] and Khadidja Henni and Brigitte Vannier and Ahmed Moussa

[1]Laboratoire de Mathématiques et Applications, UMR 7348 Univ. Poitiers et CNRS, Université de Poitiers, Poitiers, France
pierre-yves.louis@math.cnrs.fr

We present a clustering method based on sampling Marked Point Processes through an MCMC approach. The cost/Gibbs energy function is balancing different goals. (1) The data driven objective function is provided according to k-influence space. Data in a high-dense region are favored to be covered by a ball. (2) An interaction part in the energy prevents the balls full overlap phenomenon and favors connected groups of balls. Data covered through a common connected component of balls constitute then each cluster. It detects clusters of different shapes, sizes and unbalanced local densities. The curse of dimensionality is handled by using a local subspace in a weighted similarity metric. This algorithm has been applied in real benchmarks through gene expression data of various sizes. Different experiments have been done to compare this new method against the most well-known clustering algorithms. This is a joint work with Kh. Henni, B. Vannier and A. Moussa. https://hal.archives-ouvertes.fr/hal-02077905

Monday 22
16:30 - 18:30
Seminari B

Structural Breaks in Nonparametric Models via Atomic Pursuit Methods

Matus Maciak[1] and Ivan Mizera

[1]Deparment of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic
maciak@karlin.mff.cuni.cz

Changepoint estimation within the nonparametric regression framework either relies on some prior knowledge of the changepoint locations or, instead, some multistage approaches are used where one firstly detects existing changepoints and, later, the model is estimated using the knowledge gained from the first phase. We propose a direct method to changepoint detection and estimation in the nonparametric regression: a fully data-driven all-at-once approach where no additional prior knowledge for changepoint locations nor their overall number is needed.

The method combines the nonparametric regression models with the idea of sparse atomic pursuit techniques — the L1-norm regularization in particular. We discuss different model alternatives, and some theoretical properties and changepoint inference tools are derived. A finite sample performance is investigated in terms of some Monte-Carlo simulations and practical examples.

Monday 22
16:30 - 18:30
Seminari B

Non-Parametric Structural Change Detection in Multivariate Systems

Pekka Malo[1] and Lauri Viitasaari and Olga Gorskikh and Pauliina Ilmonen

[1]Information and Service Management, Aalto University, Espoo, Finland
pekka.malo@aalto.fi

Structural change detection problems are often encountered in analytics and econometrics, where the performance of a model can be significantly affected by unforeseen changes in the underlying relationships. Although these problems have a comparatively long history in statistics, the number of studies done in the context of multivariate data under nonparametric settings is still small. In this paper, we propose a consistent method for detecting multiple structural changes in a system of related regressions over a large dimensional variable space. In most applications, practitioners also do not have a priori information on the elevance of different variables, and therefore, both locations of structural changes as well as the corresponding sparse regression coefficients need to be estimated simultaneously. The method combines nonparametric energy distance minimization principle with penalized regression techniques. After showing asymptotic consistency of the model, we compare the proposed approach with competing methods in a simulation study. As an example of a large scale application, we consider structural change point detection in the context of news analytics during the recent financial crisis period.

Bootstrap Group Penalties for Predicting Molecular Properties

Tuesday 23
10:30 - 12:30
Room 8

Valentina Mameli[1] and Debora Slanzi and Irene Poli

[1]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Mestre, Venice, Italy
valentina.mameli@unive.it

The complexity of biological systems often causes scientists to generate data characterized by a huge number of variables that may affect the properties of the systems. This huge number of variables and their different nature, as well as the limited number of compositions that generally can be tested due to technical and economical constraints, often makes it difficult to apply standard statistical methodological approaches. This is certainly the case for drug discovery research, as the development of new drug therapies usually involves modelling the relations among a large set of variables describing molecules and their associated pharmacological properties and physiological effects. With the aim of modelling the complex relations between molecular properties and the very large set of variables that can affect these properties, we introduce a high dimensional regression model with a group/bi-level penalization term, extending the concepts of the individual and group variables penalties with bootstrap re-sampling techniques. In particular, the procedure first draws bootstrap samples from the original data and then selects variables and groups of variables by using a penalization procedure for each bootstrap sample. At the end of the bootstrap procedure, only variables that have been selected with a high frequency in the bootstrap replications are included in the model. The performance of the approach is shown in studying the MMP - 12 enzyme, which has a fundamental role in the development of relevant pathologies and is involved in acute and chronic pulmonary inflammatory diseases. The results suggest that the combination of penalization terms with the bootstrap procedure can provide good results when high-dimensionality characterizes the system under study.

Graph-Based Parameter-Free Clustering Algorithms Based on Ripley's K-Function

Wednesday 24
10:30 - 12:30
Room 6

Artür Manukyan[1] and Elvan Ceyhan

[1]Department of Industrial and Systems Engineering, Yeditepe University, Istanbul, Turkey
artur.manukyan@yeditepe.edu.tr

We propose clustering algorithms based on a recently developed family of geometric digraphs called cluster catch digraphs (CCD). These digraphs are used to devise clustering methods that are hybrids of density-based and graph-based clustering methods. CCDs are appealing digraphs for clustering since they estimate the number of clusters; however, CCDs, and density-based methods in general, require parameters representing the spatial intensity of assumed clusters existing in the data set. Our algorithms are parameter-free versions of the CCD algorithm that do not require specifying the spatial intensity parameter whose choice is often critical to find an optimal partitioning of the data set. We approach the problem of estimating the number of clusters by borrowing a tool from spatial data

analysis called Ripley's K-function. We call our new digraphs based on the K-function as R-CCDs. We show that the dominating sets of R-CCDs locate and separate the clusters from the noise clusters in data sets, and hence, allow us to estimate the actual number of clusters. Our parameter-free clustering algorithms are composed of methods that estimate both the number of clusters and the spatial intensity parameter, making them completely parameter-free. We conduct extensive Monte Carlo simulations and use real life experiments to compare R-CCDs with some commonly used density-based and prototype-based clustering methods and show relatively good performance of R-CCDs.

Monday 22
14:00 - 16:00
Room 7

Biases in Bias Elicitation

Giancarlo Manzi[1] and Martin Forster

[1]Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy
giancarlo.manzi@unimi.it

In recent years there has been an increasing use of expert opinion in estimation procedures, which in some cases is believed to speed the analysis more than through improved data sets. The literature in this field has proposed a number of approaches to specifically deal with bias elicitation, either in theoretical or in applied works. In this paper we contribute to this literature by considering the biases that can arise in bias elicitation when expert assessors make random errors. After presenting a general framework of the phenomenon, we illustrate it for two examples: the case of omitting variables bias and that of the bias arising in adjusting relative risks. Results show that, even when high quality assessors are tasked with making elicitations for bias, and assessors' elicitations of bias have desirable properties, the nonlinearities in biases can lead to biased elicitations of bias. For the case of omitted variables in least squares regression it is shown that the bias associated with the assessors' elicitations for point estimators will equal zero only if the elicitations have zero covariance, and that bias in bias elicitations for the variance of the point estimators is present even if the errors in the elicitations have zero covariance. In the case of bias elicitation for log relative risk as in Turner et al. (2009), if it is assumed that assessors make unbiased elicitations of the lower and upper ranges on the elicitation scale, elicitations are biased. We show the corrections, which can be made to remove this bias and discuss the implications for the applied literature, which employs these methods, also exploring the possibility of future developments for this area of research in a Bayesian framework.
References Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S., Thompson, S.G. (2009) Bias modelling in evidence synthesis. Journal of the Royal Statistical Society, Series A, 172(1): 21-47.

Discrete Statistical Models with Rational Maximum Likelihood Estimator

Tuesday 23
14:00 - 16:00
Room 5

Orlando Marigliano[1] and Eliana Duarte and Bernd Sturmfels

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
orlando.marigliano@mis.mpg.de

A discrete statistical model is a subset of a probability simplex. Its maximum likelihood estimator (MLE) is a retraction from that simplex onto the model. We characterize all models for which this retraction is a rational function. This is a contribution via real algebraic geometry which rests on results due to Huh and Kapranov on Horn uniformization. We present an algorithm for constructing models with rational MLE, and we demonstrate it on a range of instances. Our focus lies on models familiar to statisticians, like Bayesian networks, decomposable graphical models, and staged trees.

Semilinear Fractional Stochastic Differential Equations Driven by a γ-HÖlder Continuous Signal with $\gamma > 2/3$.

Wednesday 24
10:30 - 12:30
Room 10

David Márquez-Carreras[1] and Jorge A. Leon

[1]Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
davidmarquez@ub.edu

We use the techniques of fractional calculus to study the existence of a unique solution to semilinear fractional differential equation driven by a γ-Hölder continuous function θ with $\gamma \in \left(\frac{2}{3}, 1\right)$. Here, the initial condition is a function that may not be defined at zero and the involve stochastic integral with respect to θ is the extension of the Young integral given by Zähle.

Image Denoising using Corrected Information Criterion and Grouping

Tuesday 23
10:30 - 12:30
Room 10

Bastien Marquis[1] and Maarten Jansen

[1]Department of Statistics, Université libre de Bruxelles, Bruxelles, Belgium
bastien.marquis@ulb.ac.be

After the wavelet transformation of a noisy picture, the details can be viewed as variables of a high-dimensional sparse signal-plus-noise model. Finding the true nonzero details and avoiding false positives is critical in order to accurately recover the original noise free picture. In addition, an information criterion such as Mallows's Cp used as a quality rule tends to overestimate the model size, including too many falsely significant variables in order to fit the observational errors. The effects of these false positives can be tempered by

applying skrinkage on the coefficients, thus possibly introducing a substancial bias amongst large nonzeros. Without shrinkage, the false positives must be considered carefully and the expression of the information criterion corrected for the optimal balance between the sum of residual squares and the regularisation to shift towards smaller models. The correction term, whose behaviour can be described as a "mirror effect", accounts for the fact that the false positives present themselves as more significant than not selected variables whereas in reality they carry more noise than the not selected variables. The mirror effect can also apply to structured variables, although its magnitude is reduced. Combining a corrected Mallows's Cp and a grouping across a set of similarly shaped pictures can improve noise reduction amongst these pictures.

Tuesday 23
16:30 - 17:30
Second Floor

Bayesian Semi-Parametric Analysis of Multivariate Continuous Responses with Variable Selection

Benjamin Marshall[1] and Georgios Papageorgiou

[1]Statistics, Birkbeck College (University of London), London, United Kingdom
benjamin.charles.marshall@gmail.com

Performing separate regressions on multiple related response variables is time consuming, results in a less parsimonious description of the data and can lead to drawing the wrong conclusion. In this talk I will give an overview of a flexible semi-parametric model for multivariate normal responses, with automatic variable selection in both the mean and the variance models. A wide range of correlation structures are permitted and a parsimonious description of the inferred structure is made available through Dirichlet Process clustering of the correlations. I will cover the reasons for wanting to perform simultaneous regressions and demonstrate the software through a handful of case studies.

Monday 22
10:30 - 12:30
Room 6

Biclustering for Multivariate Longitudinal Data

Francesca Martella[1] and Marco Alfò and Maria Francesca Marino

[1]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Roma, Italy
francesca.martella@uniroma1.it

Applications in various domains often lead to high-dimensional data, which put up the challenge of interpreting a huge mass of data often consisting of millions of measurements. A first step towards addressing this challenge is the use of data reduction techniques, which is essential in the data mining process to reveal natural structures and to identify interesting patterns in the analyzed data. One of the most important approaches to synthesize the two modes of a high-dimensional data matrix is the joint clustering of rows and columns. This latter, named biclustering, is a data mining technique which allows for simultaneous clustering of rows and columns aiming at partitioning a data matrix into homogeneous biclusters. During the past decades biclustering approaches have been proposed by various authors in several scientific fields. For example, in the genetic field, the biclustering approach may be used to identify group of genes and an associated group of conditions over

which the genes are co-expressed. In this work, we focus on a proposal of model-based biclustering for multivariate discrete longitudinal data. These come in the form of three-way data: the first dimension identifies individuals, the second dimension identifies variables, the third one identifies time occasions. While a huge literature is devoted to clustering three-way data, the literature on biclustering three-way data is limited. Motivated by the particular case of multivariate discrete longitudinal responses, we propose a model-based biclustering approach aiming at identifying clusters of units sharing common longitudinal trajectories for subsets of variables. Specifically, a finite mixture of generalized linear models is considered to cluster units; within each mixture component, a flexible and parsimonious parameterization of the corresponding canonical parameter is adopted to identify clusters of variables evolving in a similar manner across time by making use of adequate time functions.

Compressed Monte Carlo

Tuesday 23
14:00 - 16:00
Seminari B

Luca Martino[1] and Victor Elvira

[1]Statistical Signal Processing, Universidad Rey Juan Carlos, Madrid, Spain
luca.martino@urjc.es

Bayesian models have become very popular over the last years in several fields such as signal processing, statistics and machine learning. Bayesian inference needs the approximation of complicated integrals involving the posterior distribution. For this purpose, Monte Carlo (MC) methods, such as Markov Chain Monte Carlo (MCMC) and Importance Sampling (IS) algorithms, are often employed. In this work, we introduce theory and practice of a Compressed MC (C-MC) scheme, in order to compress the information contained in a cloud of samples. CMC is particularly useful in a distributed Bayesian inference framework, when cheap and fast communications with a central processor are required. In its basic version, C-MC is strictly related to the stratification technique, a well-known method used for variance reduction purposes. Deterministic C-MC schemes are also presented, which provide very good performance. The compression problem is strictly related to moment matching approach applied in different filtering methods, often known as Gaussian quadrature rules or sigma-point methods. The connections to herding algorithms and quasi-Monte Carlo perspective are also discussed. Numerical results confirm the benefit of the introduced schemes, outperforming the corresponding benchmark methods.

Evaluating Class Effects on the Joint Student Achievements in Different Subjects: a Bivariate Semi-Parametric Mixed-Effects Model

Wednesday 24
10:30 - 12:30
Room 6

Chiara Masci[1] and Francesca Ieva and Tommaso Agasisti and Anna Maria Paganoni

[1]Department of Mathematics, Politecnico di Milano, Milano, Italy
chiara.masci@polimi.it

This work proposes an innovative statistical method to measure the impact of the class on its student achievements in multiple subjects. We propose a semi-parametric mixed-effects model with a bivariate response variable, where the random effects are assumed to follow a discrete distribution with an unknown number of support points, together with an Expectation-Maximization algorithm to estimate its parameters. The bivariate setting allows to estimate the distributions of the model coeffcients related to each response variable as well as their joint distribution. In the case study, we apply the BSPEM algorithm to data about Italian middle schools, considering students nested within classes, and we identify subpopulations of classes, standing on their effects on reading and mathematics student achievements. The proposed model is extremely informative in exploring the correlation between multiple class effects. The estimated class effects on reading and mathematics student achievements are then explained in terms of various class and school level characteristics selected by means of a LASSO regression.

Tuesday 23
10:30 - 12:30
Seminari B

The Average Conditional and Partial Kendall's Tau

Margot Matterne[1] and Irène Gijbels

[1]KU Leuven, Leuven, Belgium
margot.matterne@kuleuven.be

The dependency between two random variables can be measured by association measures like Kendall's tau, Spearman's rho... For investigating how this dependency is possibly influenced by another random variable(s) the concept of conditional association measures is useful. They measure the dependency conditional on the value of a third random variable. In order to take all possible values of the random variable on which we condition into account, the concepts of average conditional and partial association measures are useful. They give us a global measure for conditional associations. For most association measures it holds that the average conditional and the partial association measure coincide, but for Kendall's tau this is not true in general. In this talk we will try to quantify and study the differences in several settings. Since conditional association measures can be expressed as functionals of conditional copulas, the latter are a basic ingredient in the talk.

Monday 22
10:30 - 12:30
Room 7

Change-Point Detection with Multivariate Data: Two-Sample Situations

Simos Meintanis[1]

[1]Economics, National & Kapodistrian University of Athens, Athens, Greece
simosmei@econ.uoa.gr

We consider paired and two-sample break-detection procedures for vectorial observations and multivariate time series. The new methods involve L2-type criteria based on empirical characteristic functions and are easy to compute regardless of dimension. We obtain asymptotic results that allow for application of the methods to a wide range of settings involving on-line as well as retrospective circumstances with dependence between the two

time series as well as with dependence within each series. In the ensuing Monte Carlo study the new detection methods are implemented by means of resampling procedures which are properly adapted to the type of data at hand, be it independent or paired, autoregressive or GARCH structured, medium or heavy–tailed. The new methods are also applied on a real data-set from the financial sector over a time period which includes the Brexit referendum.

Finite Mixture Approximation of CARMA(p,q) Model

Thursday 25
10:30 - 12:30
Room 5

Lorenzo Mercuri[1] and Andrea Perchiazzo and Edit Rroji

[1]Università degli Studi di Milano, Milano, Italy
lorenzo.mercuri@unimi.it

In this paper, we show how to approximate the transition density of a CARMA(p,q) model driven by a time changed brownian motion based on the Laguerre polynomial. We apply this result in two situations. Firstly we derive an analytical formula for option prices when the log price follows a CARMA(p,q) model. We also propose an estimation procedure based on the approximated likelihood density. We apply our results on real data.

Extending the Validity of Frequency Domain Bootstrap Methods to General Stationary Processes

Tuesday 23
14:00 - 16:00
Room 7

Marco Meyer[1] and Efstathios Paparoditis and Jens-Peter Kreiss

[1]Inst. f. Math. Stochastik, TU Braunschweig, Braunschweig, Germany
marco.meyer@tu-bs.de

Existing frequency domain methods for bootstrapping time series have a limited range. Essentially, these procedures cover the case of linear time series with independent innovations, and some even require the time series to be Gaussian. In this paper we propose a new frequency domain bootstrap method – the hybrid periodogram bootstrap (HPB) – which is consistent for a much wider range of stationary, even nonlinear, processes and which can be applied to a large class of periodogram-based statistics. The HPB is designed to combine desirable features of different frequency domain techniques while overcoming their respective limitations. It is capable to imitate the weak dependence structure of the periodogram by invoking the concept of convolved subsampling in a novel way that is tailor-made for periodograms. We show consistency for the HPB procedure for a general class of stationary time series, ranging clearly beyond linear processes, and for spectral means and ratio statistics, on which we mainly focus. The finite sample performance of the new bootstrap procedure is illustrated via simulations.

Tuesday 23
14:00 - 16:00
Room 12

Aggregated Tests of Independence Based on HSIC Measures

Anouar Meynaoui[1] and Mélisande Albert and Béatrice Laurent and Amandine Marrel

[1]CEA Cadarache, INSA Toulouse, Saint-Paul-lez-Durance, France
anouar.meynaoui@cea.fr

Dependence measures based on reproducing kernel Hilbert spaces, also known as Hilbert-Schmidt Independence Criterion and denoted HSIC, are widely used to statistically decide whether or not two random vectors are dependent. Recently, non-parametric HSIC-based statistical tests of independence have been performed. However, these tests lead to the question of prior choice of the kernels associated to HSIC, since there is as yet no method to objectively select specific kernels. In order to avoid a particular kernel choice, we propose a new HSIC-based aggregated procedure allowing to take into account several Gaussian kernels. To achieve this, we first propose non-asymptotic single tests of level $\alpha \in (0, 1)$ and second type error controlled by $\beta \in (0, 1)$. We also provide a sharp upper bound of the uniform seperation rate of the proposed tests. Thereafter, we introduce a multiple testing procedure in the case of Gaussian kernels, considering a set of various parameters. These agregated tests are shown to be of level α and to overperform single tests in terms of uniform separation rates.

We propose two contributed talks, one after the other. The first talk from Béatrice Laurent will present a single test based on HSIC measures and its non asymptotic performances. The second talk given by Anouar Meynaoui will present the aggregated procedures, their theoretical performances and the simulation study.

The presented works are conducted within the framework of a PhD thesis funded by the CEA (French: Commissariat à l'Énergie atomique et aux Énergies alternatives). The objective is to propose new statistical methods for sensitivity analysis in support of safety studies for Sodium-cooled Fast Reactors. This methodology will be applied on a real industrial test case simulating an unprotected primary loss of flow accident.

Wednesday 24
10:30 - 12:30
Room 9

Change Point Estimation in a Dynamic Stochastic Block Model

George Michailidis[1]

[1]Statistics, University of Florida, Gainesville, United States
gmichail@ufl.edu

We consider the problem of estimating the location of a single change point in a network generated by a dynamic stochastic block model mechanism. This model produces community structure in the network that exhibits change at a single time epoch. We propose a method that comprises of the following two steps: in the first one, a least squares function is used and evaluated at each time point, but ignores the community structure and just considers a random graph generating mechanism exhibiting a change point. Once the change point is identified, in the second step, all network data before and after it are used together with a clustering algorithm to obtain the corresponding community structures and subsequently estimate the generating stochastic block model parameters. We discuss its computational complexity and identify a number of settings that the identifiability condition for successfully locating the change point is satisfied under a multitude of scenarios, including merging/splitting communities, nodes joining another community, etc. Further, we establish its rate of convergence and derive the asymptotic distribution of the change point estimators. The results are illustrated on synthetic data. Finally, we discuss technical challenges posed by coupling change point detection and community detection through clustering algorithms.

Bias in Bayes Factor and Calibrated Bayes Factor for Interval Null Hypothesis

Tuesday 23
16:30 - 17:30
Second Floor

Vishal Midya[1] and Jiangang Liao and Arthur Berg

[1]Biostatistics and Bioinformatics, Pennsylvania State University, Hershey, United States
vum41@psu.edu

This paper presents a succinct and insightful expressions for the mean of the log Bayes factor. The dependence of the Bayes factor on specified priors becomes transparent, which answers some basic questions about interpreting Bayes factor as evidence against a null or an alternative hypothesis. A straightforward resolution of the Jeffreys-Lindley paradox is derived. It also provides a powerful tool to study the behavior of the Bayes factor under different underlying distributions that may generate the data. A new concept, the neutral distribution, is proposed to quantify the bias in a log Bayes factor in favoring the null hypothesis when the data is generated under an alternative hypothesis.

Nested Filters for Joint Parameter Estimation and Tracking of State-Space Nonlinear Systems

Thursday 25
14:00 - 16:00
Room 5

Joaquín Miguez[1] and Dan Crisan and Sara Perez-Vieites

[1]Universidad Carlos III de Madrid, Leganes (Madrid), Spain
joaquin.miguez@uc3m.es

Many problems in science and engineering demand the ability to calibrate the parameters and predict the time evolution of complex dynamical systems using sequentially-collected data. Here, we introduce a general methodology for the joint estimation of the static parameters and the forecasting of the state variables of nonlinear stochastic models. The proposed scheme is Bayesian and it aims at recursively computing the sequence of joint posterior probability distributions of the unknown model parameters and its (time varying) state variables conditional on the available observations. The new framework combines two layers of inference: in the first layer, a grid-based scheme is used to approximate the posterior probability distribution of the fixed parameters; in the second layer, stochastic filtering (or data assimilation) techniques are employed to track and predict different conditional probability distributions of the state variables. Various types of procedures (deterministic grids, Monte Carlo, quasi Monte Carlo, Gaussian filters, etc.) can be plugged into both layers, leading to a wealth of algorithms. We refer to the proposed methodology as

nested filtering (NF). In the talk, we will introduce the NF methodology and highlight the key results of the underlying theory, including explicit convergence rates. Finally, we will illustrate the application of the new methods via numerical examples.

Monday 22
10:30 - 12:30
Room 7

New Consistent Characterization Based Goodness-of-Fit Tests

Bojana Milošević[1] and Marija Cuparić and Marko Obradović

[1]Faculty of Mathematics, University, Belgrade, Serbia
bojana@matf.bg.ac.rs

We present new classes of consistent goodness-of-fit tests based on V-empirical Laplace transforms. The test statistics represent the quadratic and supremum-norm distances between appropriate functions. We inspect their properties for the particular case of testing null exponential distribution. In an extensive comparison study, we show that the tests have high Bahadur efficiencies. In addition, we compare them in the cases of small and moderate sample sizes via empirical powers.

Monday 22
14:00 - 16:00
Room 6

Shape and Order Constraints in Nonparametric Regression

Alexandre Moesching[1] and Lutz Düembgen

[1]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
alexandre.moesching@stat.unibe.ch

Assuming a shape constraint in a statistical model has shown its benefit on several occasions, for example in circumstances where a parametric model is hard to justify but a shape constraint on the distribution is natural. We consider here constraints on an unknown family of distributions $(F_x)_{x \in U}$, with a fixed real subset U, and discuss nonparametric estimation procedures based on a sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ such that, conditional on the X_i, the Y_i are independent random variables with distribution functions F_{X_i}. In particular, we combine likelihood ratio ordering with a log-concavity constraint on the density f_x.

Applications of the Negative Hypergeometric Distribution

Tuesday 23
16:30 - 17:30
Second Floor

Wessel Hendrik Moolman[1]

[1]Statistics, Walter Sisulu University, Mthatha, South Africa
moolman.henri@gmail.com

The population from which the sample is drawn is the same as that for the ordinary hypergeometric distribution, but the outcome of interest is the waiting time until the kth success. The limiting distribution of the negative hypergeometric distribution can be shown to follow a negative binomial distribution. Formulae for the moments of the distribution will be derived. The following applications will be discussed. 1 Estimation of the proportion of successes in a population where the occurrence of the characteristic of interest is rare e.g. in quality control with a small proportion of defectives. 2 Estimation of M (number of success elements) when N (population size) is known. 3 Linguistic studies. 4 Study of a bonus game for slot machines.

The Impact of the International Commodity Market on the Brazilian Economy: An Analysis using Global-VAR (GVAR)

Tuesday 23
14:00 - 16:00
Room 8

George Augusto Morcerf[1] and Tarciso Da Silva and Osmani Guillen

[1]Ibmec, Rio de Janeiro, Brazil
george.morcerf@hotmail.com

This paper studies the influences of the prices of agricultural and mineral commodities on the Brazilian economy through shocks in commodities price and their impacts on GDP, exports, exchange rate and stock index. With this objective, a GVAR model composed of 57 countries was build. We found that price shocks in mineral and agricultural commodities affect GDP and exports, while for exchange rate and stock index the price shocks produce inconsistent results. Positive price shocks for both commodities generate growth in the short term with decline in the long term for GDP and exports. Negative price shocks generate converse behavior. The price shocks from agricultural commodities have more impact than mineral commodities in the Brazilian economy. As aditional result we show that the Chinese GDP has a bigger impact on international commodity prices in relation to all other countries in the study. Brazilian GDP has no impact in the international prices of the commodities considered in the study.

Thursday 25
14:00 - 16:00
Room 6

R Squared and Decision Theory

Jonas Moss[1]

[1]Department of Maths, University of Oslo, Oslo, Norway
jonas.gjertsen@gmail.com

I argue that association measures, such as R squared, should be interpreted as numerical answers to question like "How much more do I know about Y when I know X?". I show how such questions can be made precise using decision theory. Using this theory, it is possible to make association measures that have concrete interpretations, are tailored for specific situations, and make sense under any probabilistic model. I argue that some properties usually required for association measures, such as symmetry, are not desirable. Finally, I illustrate the theory with a couple of examples.

Thursday 25
10:30 - 12:30
Room 6

Optimal Solutions to the Isotonic Regression Problem

Anja Mühlemann[1] and Alexander I. Jordan and Johanna F. Ziegel

[1]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
anja.muehlemann@stat.unibe.ch

In general, the solution to a regression problem is the minimizer of a given loss criterion, and as such depends on the specified loss function. The non-parametric isotonic regression problem is special, in that optimal solutions can be found by solely specifying a functional. These solutions will then be minimizers under all loss functions simultaneously as long as the loss functions have the requested functional as the Bayes act. The functional may be set-valued. The only requirement is that it can be defined via an identification function, with examples including the expectation, quantile, and expectile functionals.

Generalizing classical results, we characterize the optimal solutions to the isotonic regression problem for such functionals in the case of totally and partially ordered explanatory variables. For total orders, we show that any solution resulting from the pool-adjacent-violators (PAV) algorithm is optimal. It is noteworthy, that simultaneous optimality is unattainable in the unimodal regression problem, despite its close connection.

Robust Sparse Covariance Estimation by Thresholding Tyler's M-Estimator

Monday 22
14:00 - 16:00
Room 12

Boaz Nadler[1] and John Goes and Gilad Lerman

[1]Weizmann Institute of Science, Rehovot, Israel
boaz.nadler@weizmann.ac.il

Estimating a high-dimensional sparse covariance matrix from a limited number of samples is a fundamental problem in contemporary data analysis. Most proposals to date, however, are not robust to outliers or heavy tails. Towards bridging this gap, we consider estimating a sparse shape matrix from n samples following a possibly heavy tailed elliptical distribution. We propose estimators based on thresholding either Tyler's M-estimator or its regularized variant. We derive bounds on the difference in spectral norm between our estimators and the shape matrix in the joint limit as the dimension p and sample size n tend to infinity with $p/n \to \gamma > 0$. These bounds are minimax rate-optimal. Results on simulated data support our theoretical analysis.

The Halfspace Depth Characterization Problem

Tuesday 23
14:00 - 16:00
Room 10

Stanislav Nagy[1]

[1]Dept. of Probability and Math. Statistics, Charles University, Prague, Czech Republic
nagy@karlin.mff.cuni.cz

The halfspace (or Tukey) depth is an inferential tool that aims to generalize quantiles to multivariate datasets. It has been long conjectured that, just as for the usual quantiles, there is a one-to-one relation between all Borel probability measures, and all possible depth surfaces. We answer this conjecture in the negative. That suggests an interesting open problem of characterizing those probability measures that possess a unique depth. A complete solution to this problem would have far-reaching implications, not only in the theory of multivariate statistics.

Missing Value Imputation in Cluster Analysis

Thursday 25
10:30 - 12:30
Seminari B

Halehsadat Nekoee Zahraei[1] and Virginie Paulus and Renaud Louis and Anne-françoise Donneau

[1]Public Health, University of Liège, Liège, Belgium
h.nekoee@uliege.be

The problem of missing values is unavoidable in clinical research. Missing value has been investigated by extensive methods. One of the attractive methods for handling missing data is multiple imputation (MI). However, applying MI method within the framework of

clustering involves several difficulties and limitations. Many clustering algorithms cannot deal with missing values, therefore, in this study, we proposed a new procedure for applying MI and variable reduction when the main goal is cluster analysis on the dataset contains missing values. The important part of this new procedure is to combine the clustering for each imputed dataset to produce the best result of the subject clustering in terms of classification. Therefore, we propose a clustering combination algorithm with a novel framework. The proposed procedure starts by applying MI technique, then uses factor analysis for reducing the complexity of high-dimensional data. In the clustering step, k-means, hierarchical and model-based are used for clustering imputed datasets. In the final merging step, we propose to use indices which have been developed to determine the optimal number of clusters to offer the best permutation of clustering for assigning the subject to the cluster. The derived results are compared with previous methods. The main difficulty in this procedure is that the cluster analysis involves many technical decisions, therefore, various algorithms are defined and compared. Simulation studies are conducted to illustrate the usefulness of our methodology against commonly used alternative models. Also, the practicality is examined by analyzing chronic obstructive pulmonary disease (COPD) that are taken from the Pneumology Department of the University hospital of Liege, which aimed to identify clinical phenotypes among adults suffering from COPD. In conclusion, our proposed procedure is very practical and flexible to allow the user to compare several methods in clustering and merging step.

Tuesday 23
16:30 - 17:30
Second Floor

Guided Structure Learning of DAGs for Count Data

Thi Kim Hue Nguyen[1] and Monica Chiogna

[1]Statistical Sciences, University of Padova, Padova, Italy
nguyen@stat.unipd.it

Current demand for modelling complex interactions between genes, combined with the greater availability of high-dimensional discrete data has led to an increased focus on structure learning for discrete data in high dimensional settings. Considering that interactions among genes can be well characterized by directed graphs where an arrow between two nodes describes a hierarchical interaction between the corresponding two genes, the problem often specializes in structure learning of Directed Acyclic Graphs (DAGs).

In the context of learning gene networks, a wealth of information is actually available, usually stored in repositories such as KEGG, about a myriad of interactions, reactions, and regulations. Such information allows to order variables following directional relationships. Here, we intend to capture this form of prior knowledge and exploit it in the development of structure learning strategies for DAGs. In particular, we assume to know the topological ordering of variables in addition to the given data. We study two new algorithms for learning the structure of DAGs, based on a modification existing algorithms, proving their theoretical consistence in the limit of infinite observations. Furthermore, we experimentally compare the proposed algorithms to a number of popular competitors, in order to study their behavior in finite samples.

Consistent Bayesian Inference for some Elliptic PDE Models

Tuesday 23
14:00 - 16:00
Room 9

Richard Nickl[1]

[1]Cambridge University, Cambridge, United Kingdom
r.nickl@statslab.cam.ac.uk

We discuss recent results about statistical inference in models arising with partial differential equations (PDEs). Specifically we will consider statistical inverse regression problems where a function u_f describes the solution of an elliptic PDE involving some unknown conductivity coefficient f, and the goal is to infer f from discrete measurements of u_f corrupted by additive Gaussian noise variables.

On some Inferences in Generalized Ornstein-Uhlenbeck Processes with Multiple Change-Points

Monday 22
16:30 - 18:30
Seminari B

Sévérien Nkurunziza[1] and Kang Fu

[1]Mathematics and Statistics, The University of Windsor, Windsor (Ontario), Canada
severien@uwindsor.ca

In this talk, we present improved inference method in generalized Ornstein-Uhlenbeck processes with multiple unknown change-points when the drift parameter satisfies uncertain constraint. A Salient feature of this investigation consists in the fact that the number of change-points and the locations of the change-points are unknown. We generalize some recent findings in five ways. First, our inference method incorporates the uncertain prior knowledge. Second, we derive the unrestricted estimator (UE) and the restricted estimator (RE) and we derive their asymptotic properties. Third, we derive a test for testing the hypothesized restriction and we derive its asymptotic power. Fourth, we propose a class of shrinkage estimators (SEs) which includes as special cases the UE, RE, and classical SEs. Fifth, we study the relative risk dominance of the proposed estimators, and we establish that SEs dominate the UE and the RE performs very well near the null hypothesis, but this performs poorly when the restriction is seriously violated. The novelty of the established results consists in the fact that the dimensions of the proposed estimators are random. Because of that, the asymptotic power of the proposed test and the asymptotic risk analysis do not follow from classical results in statistical literature.

Comparing the Impact of using Restricted Against Unrestricted Residuals in Bootstrap-Based Hypothesis Testing in a Simple Regression Model

Tuesday 23
16:30 - 17:30
Second Floor

Thobeka Nombebe[1] and Leonard Santana and James Allison

[1]School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa
tnombebe@gmail.com

This presentation investigates testing the hypothesis regarding regression coefficient parameters in a simple regression model using eight different model-based sampling methods. The eight model-based bootstrap methods for hypothesis testing that are investigated involve using either restricted (i.e. restricted under the null hypothesis) or unrestricted residuals, obtaining either restricted or unrestricted bootstrap samples, and selecting one of two different versions for the test statistic. The study seeks to explore which methods are correct or whether seemingly contradicting methods actually produce identical results. Theoretical expressions for the test statistics used in each method are obtained where it is found that some methods are indeed identical. The size and power of the tests are evaluated and compared using the results of a Monte Carlo study. There are four approaches that are found to be "good" for testing the slope regression coefficient parameter in a simple regression model. When the tests for the variance and the mean are examined, we also found that four approaches worked well, but only two of the approaches work well in both cases.

Thursday 25
14:00 - 16:00
Room 9

Berry-Esseen Bounds in the Breuer-Major CLT and Gebelein's Inequality

Ivan Nourdin[1] and Giovanni Peccati and Xiaochuan Yang

[1]RMATH, University of Luxembourg, Esch sur Alzette, Luxembourg
inourdin@gmail.com

Let us consider partial sums with dependent entries taking the form of a real-valued function evaluated in a stationary Gaussian sequence with mean 0 and variance 1. The celebrated Breuer-Major theorem asserts that, under certain conditions on the subordinating function and the covariance function of the Gaussian sequance, this sequence of partial sums converges in distribution to a normal variable after proper normalisation. In this talk, I will explain how to derive explicit Berry-Esseen bounds in the total variation distance for the Breuer-Major central limit theorem, in the case of a subordinating function satisfying minimal regularity assumptions. In particular, our results cover the case of the absolute value, that until now has been outside the scope of available techniques. Our approach is based on the combination of the Malliavin-Stein approach for normal approximations with Gebelein's inequality, bounding the covariance of functionals of Gaussian fields in terms of maximal correlation coefficients.

Monday 22
10:30 - 12:30
Room 7

Comparison of Symmetry Tests in I.i.d. and non i.i.d. Setting

Marko Obradović[1] and Blagoje Ivanović and Bojana Milošević

[1]Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
marcone@matf.bg.ac.rs

We present an extensive power study in which we compare different recent and classical symmetry tests. A special attention is given to the recent characterization based symmetry tests. We consider cases of testing symmetry around known and unknown centre in an i.i.d. setting, as well as testing symmetry of the innovations in linear and time series models.

Estimation of the Spectral Measure of Regularly Varying Random Vectors

Monday 22
14:00 - 16:00
Room 9

Marco Oesting[1] and Olivier Wintenberger

[1]Department Mathematik, Universität Siegen, Siegen, Germany
oesting@mathematik.uni-siegen.de

The extremal dependence structure of a regularly varying random vector X is fully described by its limiting spectral measure. In this talk, we investigate how to recover characteristics of the measure, such as extremal coefficients, from the extremal behaviour of convex combinations of components of X. Our considerations result in a class of new estimators of moments of the corresponding combinations for the spectral vector. We show asymptotic normality by means of a functional limit theorem and, focusing on the estimation of extremal coefficients, we verify that the minimal asymptotic variance can be achieved by an adaptive procedure.

Parameter Estimation for Misspecified Diffusion with Market Microstructure Noise

Wednesday 24
10:30 - 12:30
Room 5

Teppei Ogihara[1]

[1]University of Tokyo, Tokyo, Japan
teppei23@gmail.com

In this talk, we study parametric inference under the problems of market microstructure noise and nonsynchronous observations, which appear in high-frequency data analysis in a stock market. We study maximum-likelihood-type estimation for parametric diffusion processes with noisy, nonsynchronous observations, assuming that the true model is contained in the parametric family. We further study the case that this assumption is not satisfied. Such a model is called a misspecified model. Ogihara (2018) studied maximum-likelihood-type and Bayes-type estimation for a model of parametric diffusion processes with noisy, nonsynchronous observations, and showed asymptotic mixed normality of the estimators with the convergence rate $n^{-1/4}$. In this model, we assume that the true model is contained in the parametric family. However, in practice for high-frequency data, it is not an easy task to satisfy this assumption because several empirical facts of a stock market (intra-day seasonality, volatility clustering, complicated dependence structure of stocks, and so on) make it difficult to capture the stock microstructure. On the other hand, high-frequency data contains huge information and therefore machine learning methods such as neural network or support vector machine are useful to identify the structure of the diffusion coefficient. In this approach, we need to consider a theory of misspecified model. We will study asymptotic theory of a maximum-likelihood-type estimator for misspecified model. In this setting, the original maximum-likelihood-type estimator cannot attain the optimal convergence rate $n^{-1/4}$ due to the asymptotic bias. We construct a new estimator which attains the optimal rate by using a bias correction, and show the asymptotic mixed normality. We also introduce empirical study of high-frequency data in Tokyo Stock Exchange by using neural network.

Monday 22
16:30 - 18:30
Room 12

Local Asymptotic Normality in High-Dimensional Spiked Models

Prathapasinghe Dharmawansa, Iain M. Johnstone, and Alexei Onatski[1]

[1]University of Cambridge, Cambridge, United Kingdom
ao319@cam.ac.uk

We consider five types of high-dimensional statistical problems representing all cases in James' (1964) classification. Each of these problems rely on the generalized eigenvalues solving equation $det(H - \lambda E) = 0$, where H and E are high-dimensional Wisharts with covariance and non-centrality characteristics that depend on the problem's type. The characteristics corresponding to H and E differ by a matrix of fixed rank k whose eigenvalues (spikes) parameterize the joint distribution of the generalized eigenvalues. We show that, for the spikes located above a phase transition threshold, the asymptotic behavior of the log ratio of the joint density of the generalized eigenvalues to their joint density under a local deviation from these values depends only on the k of the largest generalized eigenvalues $\lambda_1, \ldots, \lambda_k$. Furthermore, we show that $\lambda_1, \ldots, \lambda_k$ are asymptotically jointly normal, and the statistical experiment of observing all the generalized eigenvalues converges in the Le Cam sense to a Gaussian shift experiment that depends on the asymptotic means and variances of $\lambda_1, \ldots, \lambda_k$. In particular, the best statistical inference about sufficiently large spikes in the local asymptotic regime is based on the k of the largest generalized eigenvalues only.

Tuesday 23
10:30 - 12:30
Seminari B

Experience-Based Food Insecurity Scales, a Non-Aggregative Approach to Synthesis of Indicators

Federica Onori[1] and Leonardo Salvatore Alaimo

[1]ISTAT, University of Rome La Sapienza, Rome, Italy
onori.federica@gmail.com

Food security is a complex and multidimensional phenomenon, *access to food* is one of its dimensions and experience-based food insecurity scales have been, in the last decades, the main tool for investigating it. When the phenomenon is considered at the household level, then these scales are built from a set of dichotomous items reflecting experiences related to access to food that the household possibly had to deal with in the previous months. From the point of view of the statistical treatment, these scales are mainly tackled using two different approaches: the counting approach and the Rasch model. The first one is mainly adopted at a *national level* to compute prevalences of food insecurity at different levels of severity. On the other hand, the Rasch model approach is adopted by the Food and Agriculture Organization of the United Nations (FAO) with the aim of monitoring access to food at a *global level* by producing comparable prevalences of food insecurity across countries.

Although following different statistical steps of the analysis, both the counting and the Rasch model approach consider the vector of responses of a household to a number of dichotomous items and condense all information into one value only, the range of which will determine the category of food insecurity the household belongs to. In this way, two

households with the same value for the final indicator could have potentially affirmatively answered different items, with a consequent lost of information that would instead help differantiate the two households and better suggest strategies for policy-makers.

The objective of our work is to provide, starting from the indicators of each domain of the access to food, synthesis adopting a non-aggregative approach, namely the *Partial Order Set Theory* (Poset). The resulting composite indicator, in contrast with the case of both the counting approach and the Rasch model, is not a number anymore but a Directed Acyclic Graph (DAG) called the Hasse diagram. This graph represents the set of partial comparabilities that can be established among different profiles households belong to. At the core of this approach is the idea that not all profiles resulting from answers to a set of dichotomous items can be directly and unambigously compared. Therefore, it can be of practical relvance to rely on a methodology that more realistically reflect the ordinal quality of the data.

Analysis of this work concerns data from the food security section of the "National Survey on Life Conditions" in Guatemala in 2014. The non-aggregative approach allowed us to highlight differences in the eight regions of Guatemala that would otherwise not show up if adopting an aggregative approach to the synthesis of indicators.

Convergent Estimation Algorithm for Frailty Models Based on Integrated Partial Likelihood

Thursday 25
10:30 - 12:30
Room 6

Ajmal Oodally[1] and Luc Duchateau and Estelle Kuhn

[1]MaIAGE INRA, Université Paris-Sud, Massy, France
ajmaloodally@hotmail.com

In survival analysis, we are interested in the time of occurrence of an event. The Cox model is often used in this area. It allows us to model the risk of occurrence of the event considered as a function of the baseline hazard and covariates. The regression coefficients are usually estimated by maximisation of the partial likelihood which does not depend on the baseline hazard function. Frailty models are an extension which allow to take into account heterogeneity through non observed random effects. Choosing a parametric structure for this baseline may strongly constrain the model when fitting data and lead to bad adjustment. Nonparametric approaches based on functional approximation of the baseline hazard have been proposed. We consider an alternative approach based on integrated partial likelihood following the idea of Cox to overcome the influence of the baseline hazard in the estimation procedure. We start by recalling the definitions of the integrated partial likelihood in the frailty model and the corresponding estimate. Based on this criteria, we develop a stochastic estimation algorithm to approximate its value. We prove under classical assumptions that the sequence generated by the stochastic algorithm converges almost surely toward a critical point of the integrated partial likelihood. We highlight the benefit of using the integrated partial likelihood approach for finite sample size through simulation studies and real data analysis. For instance, we show that an unsuitable choice of parametric structure on the baseline function can have a negative impact on the estimates thus showing the advantages of the integrated partial likelihood approach. We also compare the proposed estimation procedure with other estimators of the literature such as the coxme and frailtyHL R packages.

Uncertainty Analysis of the GI/M/1 Queue with Negative Customers

Thursday 25
14:00 - 16:00
Room 8

Sofiane Ouazine[1] and Karim Abbas

[1]Departement of Mathematics, Faculty of Exact Sciences, Bejaia, Algeria
wazinesofi@gmail.com

Abstract The main purpose of the current paper is to use Taylor series expansion approach for analyzing the uncertainty of the GI/M/1/N queue with negative arrivals. We then considered a non-parametric sensitivity analysis to investigate the numerical evaluation of stationary characteristics of the same queue in the case where we suppose that the negative arrival rate is not assessed in perfect manner, i.e. it is subject to propagate uncertainty. Furthermore, an approximative values of the expectation and variance of some stationary characteristics have been also obtained. This approach has the advantage of being transform-free as it avoids the use of Laplace transforms and/or numerical inversion techniques, which are predominantly used in the literature. Several numerical examples are also presented for illustrative purposes.

Mixture of Varying Coefficient Models with Random Effects

Thursday 25
14:00 - 16:00
Seminari B

Amine Ounajim[1] and Pierre-Yves Louis and Yousri Slaoui

[1]Laboratoire de mathématiques et applications, University of Poitiers, Poitiers, France
amine.ounajim@gmail.com

With technological advances, massive amounts of longitudinal data have become available which makes parametric and non-parametric modelling techniques for longitudinal data very useful. Varying coefficient models have received a lot of attention because they allow to model the links between a time varying response variable and time varying covariates using regression coefficients represented by smooth time functions which makes them very useful in the analysis of longitudinal data. Varying coefficient models are generally used without taking into account both data heterogeneity and within-subject correlation. In this paper, we propose a mixture of varying coefficient models with random effects incorporated into stochastic processes. This model allows to take into account inter-individual heterogeneity and intra-individual variations. We present a two-step procedure for estimating mixture proportions, smooth functions and covariance parameters using a modified expectation-maximization method and the backfitting algorithms. A cross-validation method is proposed for the selection of smoothing hyperparameters which impact the model performance. The proposed model is evaluated on simulated data. We show that the model allows to cluster subjects into homogenous groups and get better estimates of the model effects.

Bayesian Analysis of Hidden Markov Models

Monday 22
14:00 - 16:00
Room 10

Suleyman Ozekici[1] and Refik Soyer

[1]Industrial Engineering, Koç University, Istanbul, Turkey
sozekici@ku.edu.tr

Markov processes are used in many fields of sciences, engineering and management to represent the stochastic evolution of some random phenomena. In many interesting applications, the stochastic system under analysis is modulated by a Markov process that is hidden and not observable. However, the states of the stochastic system are either partially or fully observed. The task is to make statistical inference on the hidden Markov process based on data available on the observed states of the system. We present a computationally tractable Bayesian approach using Gibbs sampling and demonstrate it by a numerical illustration.

Persistent Homology for Kernel Density Exploration

Monday 22
10:30 - 12:30
Room 12

Tullia Padellini[1] and Pierpaolo Brutti

[1]Dipartimento di Scienze Statistiche, Sapienza University of Rome, Rome, Italy
tullia.padellini@uniroma1.it

Topology provides an increasingly popular framework for characterizing the shape of probability density functions; topological features of dimension 0 represent in fact local modes, while features of higher dimension capture circular and more complex structures in the support of the data. When densities have to be estimated by means of Kernel Density Estimators (KDE), however, the topology of the estimates may vary greatly depending on the bandwidth parameter..

Following a scale space approach we introduce the Persistence Flamelets, a new topological summary based on Persistent Homology that can be used to summarize and evaluate the topological evolution of a KDE with respect to its bandwidth, tracking the appearance and disappearance of features of arbitrary dimension. As opposed to state-of-the-art methods tackling the same problem (most noticeably SiZer and its variations), which can mainly investigate 0 dimensional topological features of 1 dimensional densities, Persistence Flamelets are "dimension-less" in the sense that they can be computed for topological features and densities of arbitrary dimension.

We show with real–data applications that the Persistence Flamelets yield similar results to SiZer when both approaches can be adopted, and we highlight the potential of the Persistence Flamelets in the case of higher dimensional features and/or densities when other tools are not available.

Monday 22
14:00 - 16:00
Room 9

Nonparametric Bayesian Estimation of the Extremal Dependence

Simone Padoan[1] and Stefano Rizzelli

[1]Bocconi University, Milano, Italy,
simone.padoan@unibocconi.it

Predicting the extremes of multiple variables is important in many applied fields for risk management. The extreme-value theory proposes several approaches for modelling multivariate extremes. According to the componentwise block-maxima approach, one possibility is to focus on the so-called max-stable models, i.e. a class of asymptotic distributions for suitable normalised componentwise maxima of random vectors.

Max-stable distributions are characherized by an extreme-value copula and margins that are members of the so-called univariate generalized extreme-value family of distributions. In particular, the extreme-value copula depends on an infinite-dimensional parameter, which is a function called the angular measure (i.e. a probability measure subjected to some constraints) that permits an interpretation of the amount of dependence.

In this contribution we discuss a fully nonparametric Bayesian estimation method for inferring the dependence of bivariate max-stable distributions based on polynomials in Bernstein form and splines. Then, we discuss the asymptotic properties of such inferential procedures. Next, we describe an extension of the first framework in arbitrary dimensions and we discuss the asymptotic properties for the resulting inferential procedure.

The asymptotic results are derived assuming that a data sample comes from a max-stable distribution with known margins. However, in practice max-stable distributions are asymptotic models, for sufficiently large sample sizes and the margins are known apart from some unknown parameters. Finally, we discuss how the asymptotic results extends to the case where the data come from a distribution that is in a neighbourhood of a max-stable distribution and to the case where the margins of the max-stable distribution are heavy-tailed with unknown tail indices.

Wednesday 24
10:30 - 12:30
Room 11

Inference for Spherical Location under High Concentration

Davy Paindaveine[1] and Thomas Verdebout

[1]Université libre de Bruxelles, Brussels, Belgium
dpaindav@ulb.ac.be

Motivated by the fact that circular or spherical data are often much concentrated around a location θ, we consider inference about θ under "high concentration" asymptotic scenarios for which the probability of any fixed spherical cap centered at θ converges to one as the sample size n diverges to infinity. Rather than restricting to Fisher-von Mises-Langevin distributions, we consider a much broader, semiparametric, class of rotationally symmetric distributions indexed by the location parameter θ, a scalar concentration parameter κ and a functional nuisance f. We determine the class of distributions for which high concentration is obtained as κ diverges to infinity. For such distributions, we then consider inference (point

estimation, confidence zone estimation, hypothesis testing) on θ in asymptotic scenarios where κ_n diverges to infinity at an arbitrary rate with the sample size n. Our asymptotic investigation reveals that, interestingly, optimal inference procedures on θ show consistency rates that depend on f. Using asymptotics "à la Le Cam", we show that the spherical mean is, at any f, a parametrically super-efficient estimator of θ and that the Watson and Wald tests for $\mathcal{H}_0 : \theta = \theta_0$ enjoy similar, non-standard, optimality properties. Our results are illustrated by Monte Carlo simulations. On a technical point of view, our asymptotic derivations require challenging expansions of rotationally symmetric functionals for large arguments of the nuisance function f.

Probabilistic Comparison of Quantile Estimators for Continuous Random Variables

Monday 22
14:00 - 16:00
Room 6

Matti Pajari[1] and Lasse Makkonen and Maria Tikanmäki

[1]Berakon, Aalto University, Espoo, Finland
matti.pajari@berakon.fi

We discuss the foundations of the criteria to assess the goodness of a quantile estimators. We provide the probabilistic justification for the Makkonen-Pajari-Tikanmäki bin criterion and show that it is a more appropriate goodness criterion than those based on minimizing the bias of the quantiles or bias of the distribution parameters. Examples of applying the bin criterion to some widely used probability distributions suggest that the Weibull plotting together with curve fitting using the nonlinear minimum squared error method is a better estimator than those applying maximum likelihood or traditional moment methods; and with a proper weighting it outperforms the maximum likelihood method as well the methods of probability weighted moments and L-moments in extreme value analysis.

Amplitude and Phase Variation of Random Processes

Wednesday 24
09:00 - 10:00
Aula Magna
Building 13

Victor M. Panaretos[1]

[1]Institute of Mathematics, EPFL, Switzerland
victor.panaretos@epfl.ch

The amplitude variation of a random process consists of random oscillations in its range space (the “y-axis”), typically encapsulated by its (co)variation around a mean level. In contrast, phase variation refers to fluctuations in its domain (the “x-axis”), often caused by random time changes or spatial deformations. Many types of processes, particularly physiological processes, manifest both types of variation, and confounding them can seriously skew statistical inferences. We will consider some of the statistical challenges related to empirically separating these two forms of variation, a problem also known as registration, synchronisation, or multireference alignment, in other contexts. Our approach will largely rely on the tools and geometry of optimal (multi)transport, and borrow from connections to notions from shape theory, such as Procrustes analysis and tangent space PCA. The approach will hopefully also highlight the intriguing aspect of this problem, as being at

the confluence of functional data analysis, where the data are elements of infinite dimensional vector spaces, and geometrical statistics, where the data are elements of differentiable manifolds.

Tuesday 23
14:00 - 16:00
Room 7

Recent Advances in Bootstrapping Functional Time Series

Efstathios Paparoditis[1]

[1]Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus
stathisp@ucy.ac.cy

We review some recent developments in bootstrapping functional time series and focus on applications to testing equality of interesting characteristics for several stationary functional processes using fully functional test statistics.

Tuesday 23
16:30 - 17:30
Second Floor

Parameter Estimation for Non-Stationary Fisher-Snedecor Diffusion

Ivan Papić[1] and Alexey M. Kulik and Nikolai N. Leonenko and Nenad Šuvak

[1]Department of Mathematics, J. J. Strossmayer University of Osijek, Osijek, Croatia
ipapic@mathos.hr

The problem of parameter estimation for the non-stationary ergodic diffusion with Fisher-Snedecor invariant distribution, to be called Fisher-Snedecor diffusion, is considered. We propose generalized method of moments (GMM) estimator of unknown parameter, based on continuous-time observations, and prove its consistency and asymptotic normality. The explicit form of the asymptotic covariance matrix in asymptotic normality framework is calculated according to the new iterative technique based on evolutionary equations for the point-wise covariations. The results are illustrated in a simulation study covering various starting distributions and parameter values.

Monday 22
14:00 - 16:00
Room 6

Testing for Superiority Between Two Variance Functions

Juan Carlos Pardo-Fernandez[1] and Graciela Boente

[1]Departament of Statistics and Operations Research, Universidade de Vigo, Vigo, Spain
juancp@uvigo.es

This paper focuses on the problem of testing the null hypothesis that the variance functions of two populations are equal versus one-sided alternatives under a general nonparametric heteroscedastic regression model. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root-n contiguous alternatives. A Monte Carlo study is performed to analyse the finite sample behaviour of the proposed test.

Tests on the Block-Diagonal Covariance Matrix with a Large Number of Blocks

Wednesday 24
10:30 - 12:30
Room 10

Nestor Parolya[1] and Holger Dette and Taras Bodnar

[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands
n.parolya@tudelft.nl

In this paper we extend the previous tests for testing the block-diagonality of large covariance matrix to an arbitrary number of blocks. We consider the case where the dimension of the vectors increases with the sample size and propose multivariate analysis of variance-type statistics for the hypothesis of a block diagonal covariance matrix. Furthermore, we investigate the case when the number of blocks tends to infinity and provide the asypmtotic distribution of the test statistics in this situation. The asymptotic properties are investigated under both the null hypothesis and the alternative hypothesis using random matrix theory. For this purpose we study the weak convergence of linear spectral statistics of central and (conditionally) non-central Fisher matrices. The case with infinite number of blocks corresponds to infinite number of linear spectral statistics, which is not considered in the literature before.

On AFT Mixture Cure Models, Benefits and Estimation

Friday 26
09:00 - 11:00
Room 6

Motahareh Parsa[1] and Ingrid Van Keilegom

[1]ORSTAT, KU Leuven, Heverlee, Iran
motahareh.parsa@kuleuven.be

When studying survival data in the presence of right censoring, it is common to face that a certain proportion of the individuals under study do not experience the event of interest and are considered as cured ones. The mixture cure model is a generalized form of classical survival model, and is a regression model that takes the cure fraction into account, it depends on a model for the conditional probability of being cured (called the incidence) and a model for the conditional survival function of the uncured individuals (called the latency). This work considers a logistic model for the incidence and a semiparametric accelerated failure time (AFT) model for the latency part. The estimation of this model is obtained via maximizing the likelihood, in which the unknown error density is replaced by a kernel estimator based on the Kaplan-Meier estimator of the error distribution. Necessary theories for consistency and asymptotic normality of the parameter estimators are provided. Nonetheless, the introduced estimation method is compared with some proposed method such as given in Lu (2010), which uses a kernel approach based on the EM algorithm to estimate the model parameters. Ultimately, the aforementioned method is applied on a cancer data set and results are demonstrated.

Tuesday 23
14:00 - 16:00
Room 8

A Viral Approach to Early Prediction of Adoptions of New Products

Riccardo Parviero[1] and Ida Scheel and Kristoffer Hellton and Geoffrey Canright

[1]Department of Mathematics, Universitetet i Oslo, Oslo, Norway
riccarpa@math.uio.no

One of the main questions that arise upon the launch of a new product is to predict how popular it can become. Assuming that the customers' relationships can be represented as a network, we propose a framework which estimates the parameters that govern the adoption spread and then performs predictions through simulations of the evolution of the adoption process on out-of-sample data. Our method can capture both time-varying external influences, e.g. advertisement or interest, and viral influences, when peer-to-peer interactions lead the adoption spread. This method interprets product adoptions as events drawn from a trajectory of an inhomogeneous Poisson process in which the rate is modified by the covariates of a specific individual and its neighbours. In this framework, every customer is influenced by (and, conversely, may influence) its immediate neighbours, which, in the network representation, are the individuals they share an edge with. An edge between two people is here defined to exist if they have shared an amount of phone calls or text messages which exceeds a certain threshold. We tested our method on adoptions of a real product and on synthetic adoptions. In both cases the underlying customers' network is built using real telephone data of a European country. These two tests show that the method is reliable in every setting, even when either one of its components dominates on the other. We also investigate how many adoptions of the product our model needs to start producing accurate estimates. Surprisingly, our method converges to stable estimates, thus producing reliable predictions, after a small number of adoptions (roughly the first 10

Tuesday 23
14:00 - 16:00
Room 10

Rate-Optimal Estimators for the Volume of a Set

Beatriz Pateiro-López[1] and Ery Arias-Castro and Alberto Rodríguez-Casal

[1]Estatística, Análise Matemática e Optimización, Santiago de Compostela, Santiago de Compostela, Spain
beatriz.pateiro@usc.es

In this work, our goal is to estimate the volume of a compact domain based on a sample from the uniform distribution on that domain. We restrict the class of sets under consideration to domains with boundary with positive reach. The family of sets that fulfil this condition is much wider than that of convex sets, where the problem of volume estimation has already been considered. Plug-in type estimators are a natural choice for the estimation of functionals of a set such as the volume. For instance, in the convex case, the volume of the convex hull of the sample is a natural estimator for the volume of the set. However, the plug-in estimator based on the sample convex hull is not rate-optimal for the estimation of the volume. In our setting, when the compact domain is assumed to be r-convex, a natural choice would be to estimate the set by the so-called r-convex hull of the sample points. Again, the plug-in estimator based on the sample r-convex hull is not rate-optimal for the estimation of the volume. Thus, we propose an estimator based on the r-convex hull of the sample that applies a simple correction of the bias based on sample splitting. We show

that this simple estimator achieves a minimax lower bound that we derive. Our method of estimation can aso be used to provide a confidence interval for the volume of the set.

Markov Properties of the Common Factor Analytic Model

Wednesday 24
10:30 - 12:30
Room 6

Carel Peeters[1]

[1]Epidemiology & Biostatistics, Amsterdam University medical centers, Location VUmc, Amsterdam, Netherlands
cf.peeters@vumc.nl

The factor analytic model assumes that a random vector of correlated variables can be projected onto a lower-dimensional latent space. This model has a rich tradition in many fields, such as psychometrics, statistics, and machine learning. It is used both in an exploratory mode as a data reduction technique and in a confirmatory mode as a generative description of the data. In this presentation a complete Markov characterization of the common (exploratory and confirmatory) factor analytic model is given. This enables one to establish conceptual bridges between (psychometric) path diagrams and directed bipartite mixed graphs. In doing so, some implications for factor analytic modeling are discussed, especially with respect to the determinacy of the latent projection when the number of variables grows to infinity. It is then shown how these new insights can be of use in the analysis of data stemming from the high-throughput mining of quantitative features from medical images.

Estimation and Clustering in Popularity Adjusted Stochastic Block Model

Thursday 25
10:30 - 12:30
Room 11

Marianna Pensky[1] and Majid Noroozi and Ramchandra Rimal

[1]Department of Mathematics, University of Central Florida, Orlando, United States
marianna.pensky@ucf.edu

The paper considers the Popularity Adjusted Block model (PABM) introduced by Sengupta and Chen (2018). We argue that the main appeal of the PABM is the flexibility of the spectral properties of the graph which makes the PABM an attractive choice for modeling networks that appear in biological sciences. We expand the theory of PABM to the case of an arbitrary number of communities which possibly grows with a number of nodes in the network and is not assumed to be known. We produce the estimators of the probability matrix and the community structure and provide non-asymptotic upper bounds for the estimation and the clustering errors. We use the Sparse Subspace Clustering (SSC) approach to partition the network into communities, the approach that, to the best of our knowledge, has not been used for clustering network data. The theory is supplemented by a simulation study. In addition, we show advantages of the PABM for modeling a butterfly similarity network and a human brain functional network.

Monday 22
16:30 - 18:30
Room 6

Asymptotically Optimal Pointwise and Minimax Changepoint Detection for General Stochastic Models with a Composite Post-Change Hypothesis

Serguei Pergamenchtchikov[1] and Alexander Tartakovsky

[1]Laboratory of Mathematics Raphaël Salem, University of Rouen,
Saint-Étienne-du-Rouvray, France
Serge.Pergamenchtchikov@univ-rouen.fr

A weighted Shiryaev-Roberts change detection procedure is shown to approximately minimize the expected delay to detection as well as higher moments of the detection delay among all change-point detection procedures with the given low maximal local probability of a false alarm within a window of a fixed length in pointwise and minimax settings for general non-i.i.d. data models and for the composite post-change hypothesis when the post-change parameter is unknown. We establish very general conditions for the models under which the weighted Shiryaev–Roberts procedure is asymptotically optimal. These conditions are formulated in terms of the rate of convergence in the strong law of large numbers for the log-likelihood ratios between the "change" and "no-change" hypotheses, and we also provide sufficient conditions for a large class of ergodic Markov processes. Examples related to multivariate Markov models where these conditions hold are given.

Thursday 25
14:00 - 16:00
Room 8

Numerical Evaluation of the Transition Probability of the Simple Birth-and-Death Process

Alberto Pessia[1] and Jing Tang

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
alberto.pessia@helsinki.fi

The simple birth-and-death process is a continuous-time Markov process that is commonly employed for describing changes over time of the size of a population. Application areas are for example in queueing theory, demography, epidemiology, or biology. Although the transition probability is available in closed form, its direct evaluation can be numerically unstable and prevent the use of the birth-and-death process in real settings, for example for maximum likelihood estimation. I will show under which conditions the transition probability can be numerically unstable and present an alternative representation in terms of the Gaussian hypergeometric function. I will also show how the hypergeometric function can be evaluated efficiently and accurately using a three-term recursive relation. I will finally give an example on how to use the hypergeometric representation to numerically find the maximum likelihood estimator.

Changepoint in Non-Stationary Series Without Nuisance Parameters

Monday 22
16:30 - 18:30
Seminari B

Michal Pesta[1] and Martin Wendler

[1]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
michal.pesta@mff.cuni.cz

Many changepoint detection procedures rely on the estimation of nuisance parameters (like long-run variance). If a change has occurred, estimators might be biased and data-adaptive rules for the choice of tuning parameters might not work as expected. If the data is not stationary, this becomes more challenging. The aim of this paper is to present two changepoint tests, which involve neither nuisance nor tuning parameters. This is achieved by combing self-normalization and wild bootstrap. We investigate the asymptotic behavior and show the consistency of the bootstrap under the hypothesis as well as under the alternative, assuming mild conditions on the weak dependence of the time series. As a by-product, a changepoint estimator is introduced and its consistency is proved. The results are illustrated through a simulation study. The new completely data-driven tests are applied to a real data example from finance.

The Impossibility of Conditional Independence Testing and a Way Out

Monday 22
16:30 - 18:30
Room 9

Jonas Peters[1] and Rajen Shah

[1]University of Copenhagen, Copenhagen, Denmark
jonas.peters@math.ku.dk

Several statistical procedures rely on testing for conditional independence, i.e., on testing whether two random variables X and Y carry information about each other, given the observation of a third variable Z. We prove that, statistically speaking, testing for conditional independence is a fundamentally hard problem if the conditioning variable Z is continuous. Solving it requires carefully chosen assumptions on the data generating process. We propose a practical test that achieves the correct size if the conditional expectations are smooth enough such that they can be estimated from data. This is joint work with Rajen Shah.

Componentwise Estimation of Ordered Scale Parameters of Two Exponential Distributions under a General Class of Loss Function

Wednesday 24
10:30 - 12:30
Room 6

Constantinos Petropoulos[1] and Lakshmi Kanta Patra and Somesh Kumar

[1]Department of Mathematics, University of Patras, Patras, Greece
costas@math.upatras.gr

In many real life situations, prior information about the parameters is available, such as the order of the parameters. Incorporating this prior information about the order restrictions on parameters leads to more efficient estimators. In the present communication, we investigate estimation of the ordered scale parameters of two shifted exponential distributions with unknown location parameters under a general class of scale invariant loss functions. It is proved that the best affine equivariant estimator (BAEE) is inadmissible. We have obtained various non smooth and smooth estimators which improve upon the BAEE. In particular we have derived the improved estimators for special loss function.

Monday 22
16:30 - 18:30
Room 9

Kernel-Based Tests for Joint Independence

Niklas Pfister[1] and Peter Bühlmann and Bernhard Schölkopf and Jonas Peters

[1]ETH Zurich, Zurich, Switzerland
niklas.pfister@stat.math.ethz.ch

In this talk, we introduce a method for testing whether d (possibly multivariate) random variables, are jointly (or mutually) independent. Our method builds on ideas from the two-variable Hilbert–Schmidt independence criterion but allows for an arbitrary number of variables. We embed both the joint distribution and the product distribution of the marginals in a reproducing kernel Hilbert space and define the d-variable Hilbert–Schmidt independence criterion dHSIC as the squared distance between these embeddings. In the population case, the value of dHSIC is 0 if and only if the d variables are jointly independent, as long as the kernel is characteristic. While hypothesis tests built on dHSIC have some nice asymptotic properties, we will discuss some difficulties that arise when proving the “gold standard” asymptotic result: uniform asymptotic level and pointwise asymptotic power.

Tuesday 23
16:30 - 17:30
Second Floor

Hierarchical Spatial Survival Models with Application to Life-Cell Imaging Data

Thi Huong Phan[1] and Giuliana Cortese

[1]Department of Information Engineering, University of Padova, Padova, Italy
thihuong.phan@studenti.unipd.it

Exocytosis on beta-cells is one of the fundamental cellular processes that releases insulin-containing secretory granules to blood through the plasma membrane due to stimulus. Studying survival of granules on the plasma membrane and their spatial correlation within cells during the exocytosis is of great interest to researchers in biological and medical area, as it is closely related to the regulation of insulin level in blood. Data are a collection of TIRF images recorded from 8 human beta-cells, containing granules and syntaxin information. One of the main objectives of this thesis is to investigate the relationship between the survival rates of granules and syntaxin levels, while adjusting for spatial correlation among granules within cells. To answer our specific biological problem, we propose a semi-parametric proportional hazard model, where the baseline hazard function is estimated

nonparametrically and a multivariate normal distribution is assumed for individual frailties. Hence, the clustering structure, as well as the spatial correlation between granules are modeled via the variance-covariance matrix of frailties. We firstly extend the penalized partial likelihood method and the Monte-Carlo EM method to estimate the parameters in the model. Then, we contribute a novel inferential approach based on pairwise likelihood, EM algorithm and quadrature approximation. We conduct simulations to validate and compare three approaches, hence the advantages and disadvantages for each approach are discussed. Finally, we apply our method to the exocytosis data and interpret the results.

Deciding About the Emptiness of the Interior of a Manifold Based on a Dependent Sample of Points

Wednesday 24
10:30 - 12:30
Room 8

Nuno Picado[1] and Paulo Eduardo Oliveira

[1]University of Coimbra, Coimbra, Portugal
nunopicado@sapo.pt

Let M be a compact manifold of a d dimensional euclidean space. Our goal is to decide, based on a sample of points with a certain type of dependence, whether the interior of M is empty or not. We will consider two diferent models: the noiseless model, where the sample is drawn from a distribution with support on M; and the noisy model where the sample comes from a distribution with support on the set of points where the distance to M is less than a fixed constant. In both cases, we will state the asymptotic convergence of a test that decides if the interior of M is empty or not. This test is based on an estimator introduced by Devroye and Wise (1980), which uses balls centered on the sample points. The main problem is to choose the radius of these balls, in order to achieve the correct decision with a large probability.

Measuring Reliability of Functional Data: An Application to Human Movements

Monday 22
10:30 - 12:30
Room 10

Alessia Pini[1] and Jonas Markström and Lina Schelin

[1]Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy
alessia.pini@unicatt.it

Advanced measurement systems are often used in biomechanics to capture human motion during tasks, with the aim of understanding the characteristics of motion and assessing differences between groups of subjects. A common feature of such systems is that they measure functional data. Human movement has an inherent natural variation, and it is obvious that we cannot expect observed movement functional data to be identical when a task is repeated. Still, it is crucial that measurement tools are reliable, i.e., that the measures of a quantity on the same subject are consistent when repeating a task. Many functional data analysis methods have been used to analyze human movements data, for performing clustering, alignment, inference (among others). However, the literature regarding reliability of curve data is surprisingly overlooked, and reliability is often assessed on

univariate or multivariate data extracted from the original functional data. A few works on reliability for functional data have been proposed in the literature of reliability and biomechanics, but with limitations and no clear recommendations. In this talk, we identify a set of methods proposed in the biomechanics literature for measuring reliability of functional data, together with methods from functional data analysis that could be used for the same scope. We identify the positive and negative characteristics of each method and highlight their differences. Further, all methods are compared both on simulated data and on an application to knee kinematic data. Finally, possible directions for performing inference on reliability of functional data are discussed.

Tuesday 23
10:30 - 12:30
Seminari B

Community Detection for Probabilistic Graphical Models

Eugen Pircalabelu[1] and Gerda Claeskens

[1]ISBA, UC Louvain, Louvain-la-Neuve, Belgium
eugen.pircalabelu@uclouvain.be

A new strategy of probabilistic graphical modeling is developed that draws parallels from social network analysis. Probabilistic graphical modeling summarizes the information coming from multivariate data in a graphical format where nodes, corresponding to random variables, are linked by edges that indicate dependence relations between the nodes. The purpose is to estimate the structure of the graph (which nodes connect to which other nodes) when data at the nodes are available. On the opposite side of the spectrum, social network analysis considers the graph as the observed data. Given thus the graph where connections between nodes are observed rather than estimated, social network analysis estimates models that represent well an underlying mechanism which has generated the observed graph.

We propose a new method that exploits the strong points of each framework as it estimates jointly an undirected graph and communities of homogenous nodes, such that the structure of the communities is taken into account when estimating the graph and conversely, the structure of the graph is accounted for when estimating homogeneous communities of nodes. The procedure uses a joint group graphical lasso approach with community detection-based grouping, such that some groups of edges co-occur in the estimated graph. The grouping structure is unknown and is estimated based on community detection algorithms.

Theoretical derivations regarding graph convergence and sparsistency, as well as accuracy of community recovery are included, while the method's empirical performance is illustrated in an fMRI context, as well as with simulated examples.

On Optimal Estimation of Random Quantities Associated with Levy Processes

Tuesday 23
14:00 - 16:00
Room 6

Mark Podolskij[1] and Jevgenijs Ivanovs

[1]Mathematics, Aarhus, Denmark
mpodolskij@math.au.dk

In this talk we will present new results on optimal estimation of certain random quantities associated with Levy processes. More specifically, our goal is to construct efficient estimators of supremum and local time of Levy processes given high frequency observations. We consider the L^2 and L^1 losses as criteria functions. The methodology is based upon recent asymptotic results on zooming-in of a Levy process at its supremum.

Bump Detection in the Presence of Dependency

Thursday 25
10:30 - 12:30
Room 6

Markus Pohlmann[1] and Frank Werner and Farida Enikeeva and Axel Munk

[1]Institute for Mathematical Stochastics, University of Göttingen, Göttingen, Germany
markus.pohlmann@stud.uni-goettingen.de

We study the problem of detecting a bump in a dependent Gaussian processes, i.e. an abrupt change in the mean function of the process within a certain interval. We will provide an asymptotically sharp minimax detection boundary in terms of the asymptotic behaviors of the bump's length and height as well as the dependency structure of the process. Our proofs are based on laws of large numbers for non-independent and non-identically distributed arrays of random variables. A major finding is that generically positive correlation makes detection harder whereas negative correlation eases bump detection.

Predictive Inference for Locally Stationary Time Series

Tuesday 23
14:00 - 16:00
Room 7

Dimitris Politis[1] and Srinjoy Das

[1]Mathematics, Univ. of California at San Diego, La Jolla, United States
dpolitis@ucsd.edu

The Model-free Prediction Principle of Politis (2015) has been successfully applied to general regression problems, as well as problems involving stationary time series. However, with long time series, e.g. annual temperature measurements spanning over 100 years or daily nancial returns spanning several years, it may be unrealistic to assume stationarity throughout the span of the dataset. We will show how Model-free Prediction can be applied to handle time series that are only locally stationary, i.e., they can be assumed to be stationary only over short time-windows. Both one-step-ahead point predictors and prediction intervals are constructed, and the performance of model-free is compared to model-based prediction using models that incorporate a trend and/or heteroscedasticity. Both aspects of the paper, model-free and model-based, are novel in the context of time-series that are locally (but not globally) stationary. [This is joint work with Dr. Srinjoy Das]

Thursday 25
14:00 - 16:00
Room 12

Confusion: Developing an Information-Theoretic Secure Approach for Multiple Parties to Pool and Unify Statistical Data, Distributions and Inferences.

Murray Pollock[1] and Hongsheng Dai and Gareth Roberts and Louis Aslett

[1]Statistics, University of Warwick, Coventry, United Kingdom
m.pollock@warwick.ac.uk

"Monte Carlo Fusion" (Dai, Pollock, Roberts, 2018, JAP) is a new theory providing a framework for the unification of distributed statistical analyses and inferences, into a single coherent inference. This problem arises in many settings (for instance, expert elicitation, multi-view learning, distributed 'big data' problems etc.). Monte Carlo Fusion is the first general statistical approach which avoids any form of approximation error in obtaining the unified inference, and so has broad applicability across a number of statistical applications.

A direction of particular interest for broad societal impact is in Statistical Cryptography. Considering the setting in which multiple (potentially untrusted) parties wish to securely share distributional information (for instance in insurance, banking and social media settings), Fusion methodology offers the possibility that distributional sharing can be conducted in such a manner that the information which is required to be exchanged between the parties for the methodology can be secretly shared. As a consequence a gold-standard information theoretic security of the raw data can be achieved. So called "Confusion", a confidential fusion approach to statistical secret sharing, has the property that another party with unbounded compute power could not determine secret information of any other party.

Joint work with Louis Aslett, Hongsheng Dai, Gareth Roberts.

Thursday 25
10:30 - 12:30
Room 8

A Framework for Adaptive MCMC Targeting Multimodal Distributions

Emilia Pompe[1] and Chris Holmes and Krzysztof Latuszynski

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom
emilia.pompe@stats.ox.ac.uk

We propose a new Monte Carlo method for sampling from multimodal distributions (Jumping Adaptive Multimodal Sampler). The idea of this technique is based on splitting the task into two: finding the modes of the target distribution and sampling, given the knowledge of the locations of the modes. The sampling algorithm is based on steps of two types: local ones, preserving the mode, and jumps to a region associated with a different mode. Besides, the method learns the optimal parameters while it runs, without requiring user intervention. Our technique should be considered as a flexible framework, in which the design of moves can follow various strategies known from the broad MCMC literature.

In order to design an adaptive scheme that facilitates both local and jump moves, we introduce an auxiliary variable representing each mode and we define a new target distribution on an augmented state space. As the algorithm runs and updates its parameters, the target distribution also keeps being modified. This motivates a new class of algorithms, Auxiliary Variable Adaptive MCMC. We prove general ergodic results for the whole class before specialising to the case of our algorithm.

The main properties of our method will be discussed and its performance will be illustrated with several examples of multimodal target distributions.

A Bivariate Binomial Count Time Series with Application to the Number of Rainfall Days

Monday 22
16:30 - 18:30
Room 8

Božidar Popović[1]

[1]Faculty of Science and Mathematics, University of Montenegro, Podgorica, Montenegro
bozidarpopovic@gmail.com

We introduce a new bivariate count (INAR) time series without using any information about the bivariate distribution of the bivariate time series. Some properties such as strictly stationarity and ergodicity are studied. We study correlation structure and different estimation procedures such as Yule - Walker and conditional maximum likelihood. The performance of studied estimators is considered via standard deviations. This model is successfully applied in modelling number of rainfall days in Montenegro.

Detection of Changes in Panel Data Models with Stationary Regressors

Monday 22
10:30 - 12:30
Room 7

Charl Pretorius[1] and Marie Hušková and Adam Láf

[1]Charles University, Prague, Czech Republic
pretorius@karlin.mff.cuni.cz

We consider a panel regression model with cross-sectional dimension N. The aim is to test, based on T observations, whether the intercepts in the model remain unchanged throughout the observation period. The test procedure involves the use of a CUSUM-type statistic derived from a pseudo-likelihood argument. We present asymptotic results of the test statistic and break point estimators in the case where both N and T are allowed to become large. The asymptotic results are valid under strong mixing and stationarity assumptions on the error and regressor sequences. Monte Carlo results will be presented that indicate that the tests work in the case of small to moderate sample sizes. The talk ends with an illustrative application of the procedure to financial data.

Friday 26
09:00 - 11:00
Room 9

Tests for Qualitative Features in the Random Coefficients Model

Katharina Proksch[1] and Fabian Dunker and Konstantin Eckle and Johannes Schmidt-Hieber

[1]Department of Applied Mathematics, University of Twente, Enschede, The Netherlands
katharina.proksch@rub.de

The random coefficients model is an extension of the linear regression model that allows for unobserved heterogeneity in the population by modeling the regression coefficients as random variables. Given data from this model, the statistical challenge is to recover information about the joint density of the random coefficients which is a multivariate and ill-posed problem. Because of the curse of dimensionality and the ill-posedness, pointwise nonparametric estimation of the joint density is difficult and suffers from slow convergence rates. Larger features, such as an increase of the density along some direction or a well-accentuated mode can, however, be much easier detected from data by means of statistical tests. In this talk, we follow this strategy and construct tests and confidence statements for qualitative features of the joint density, such as increases, decreases and modes. We propose a multiple testing approach based on aggregating single tests which are designed to extract shape information on fixed scales and directions. Using recent tools for Gaussian approximations of multivariate empirical processes, we derive expressions for the critical value. We apply our method to simulated and real data.

Tuesday 23
14:00 - 16:00
Room 6

Asymptotic Normality of Integrated Periodogram Operators

Daniel Rademacher[1] and Jens-Peter Kreiß and Efstathios Paparoditis

[1]Institute for Mathematical Stochastics, Braunschweig, Germany
daniel.rademacher@tu-bs.de

Consider a strictly stationary (functional) process $(X_t)_{t\in\mathbb{Z}}$ taking values in $L^2([0,1])$. A key element of a frequency domain framework for drawing statistical inference on the second-order structure of the process is the spectral density operator, which generalises the notion of a spectral denisty matrix to the functional setting. As an integral operator, the spectral denisty operator is fully determined by its corresponding kernel, which can be estimated by a smoothed version of the periodogram kernel (the functional analogue to the periodogram matrix). More general many relevant quantities of the process such as autocovariance operators or spectral distribution operators can be represented as a weighted Bochner integal of the spectral density kernel. Estimators for such quantities are obtained by replacing the spectral density kernel with the periodogram kernel. Thus the class of *integrated periodogram operators* covers many familiar statistics, including empirical autocovariance and smoothed periodogram operators. We show that any finite collection of such estimators converges to a collection of jointly complex normally distributed operators. As a side-result we obtain joint asymptotic normality for the empirical autocovariance operators. Another by product is the complex normality for the spectral distribution operator. Our results do not depend on structural modeling assumptions, but only on functional versions of cumulant mixing conditions. In order to formulate these conditions precisely, we also introduce

a general form of cumulants for random elements taking values in an arbitrary separable Hilbert space.

Recursive Copula Additive Models to Estimate the Effect of a Binary Endogenous Variable in a Count Regression

Wednesday 24
10:30 - 12:30
Seminari B

Rosalba Radice[1] and Giampiero Marra and David Zimmer

[1]Cass Business School, City, University of London, London, United Kingdom
radice.rosalba@city.ac.uk

This paper present a general estimation approach, based on copula functions, for consistently estimating the coefficient of an endogenous regressor in nonlinear settings. The proposed modeling framework is evaluated in a simulation study and illustrated using a real application. Both the simulation exercises and the real-data application offer evidence that the proposed method is capable of more consistently estimating endogeneity effects relative to more naive methods. The case study estimates the effect of insurance status (a binary variable) on doctor visits (a count variable). The method finds statistically significant evidence that insurance is endogenous with respect to usage of doctor services. When that endogeneity is taken into account, the effect of insurance is larger than when endogeneity is ignored. The relevant numerical computations can be easily carried out using the freely available GJRM R package.

Machine Learning in Finance

Thursday 25
14:00 - 16:00
Seminari B

Dragana Radojicic[1]

[1]Financial and Actuarial Mathematics (FAM), Vienna University of Technology, Vienna, Austria
gagaradojicic@gmail.com

Machine learning is becoming present in various fields, and quantitative trading is one of the possible application of its concept. The goal is to analyze the informativeness of the limit order book. This research is based on a high-quality limit order book data, more precisely on the data set from the Nasdaq Stock Market (second largest exchange in the world). The data set has been aggregated the data set with respect to the different time intervals and prepared it into the desired form for research. Furthermore, the technical indicators are employed and concatenated into the data set. The concept is enhanced by extracting the features which are relevant for this research, using the genetic algorithms.

Tuesday 23
16:30 - 17:30
Second Floor

Exact Parametric Causal Mediation Analysis for Binary Outcomes with Binary Mediators

Martina Raggi[1] and Marco Doretti and Elena Stanghellini

[1]Information Management Institute, University of Neuchâtel, Neuchâtel, Switzerland
martina.raggi@unine.ch

For a binary outcome, we derive the exact parametric expressions of natural direct and indirect effects, on the odds-ratio scale, in settings with a binary mediator. The effect decomposition we propose does not require the outcome be rare and generalizes the existing ones (Gaynor et al., 2018; Samoilenko et al., 2018; Valeri and VanderWeele, 2013), allowing for interactions between both the exposure and the mediator and confounding covariates and by providing formulas for standard errors based on the delta method. Further, it outlines a more interpretable relationship between the causal effects and the correspondent pathway-specific logistic regression parameters. Our findings are applied to data from a randomized experiment in microfinance performed in Bosnia and Herzegovina (Augsburg et al., 2015). Specifically, we assess whether – and to what extent - the effect of randomly allocated microcredit loans on clients' "bankability" (that is, the capability to attract credit from financial institutions) is mediated by owning a business. Moreover, we assess the efficiency of our expressions through a simulation study.

Thursday 25
10:30 - 12:30
Room 8

Doubly Stochastic Exponential Pulse Models for Rainfall

Nadarajah Ramesh[1] and Gayatri Rode

[1]School of Computing and Mathematical Sciences, University of Greenwich, Maritime Greenwich Campus, Greenwich, London, SE10 9LS, United Kingdom
n.i.ramesh@gre.ac.uk

Doubly stochastic exponential pulse models for rainfall

Nadarajah Ramesh and Gayatri Rode School of Computing and Mathematical Sciences University of Greenwich Maritime Greenwich Campus Old Royal Naval College, Park Row, Greenwich London SE10 9LS, UK N.I.Ramesh@greenwich.ac.uk

Stochastic point process models have been used extensively to model rainfall data collected in various format. Doubly stochastic Poisson processes provide a rich class of models for analysing rainfall time series. One class of model arises when the underlying stochastic process becomes a continuous-time irreducible Markov process X(t) on a finite state space. Models from this class have been used by several authors to describe temporal rainfall characteristics at different levels of aggregations. Ramesh et al. (2012, 2017) described this class of models for analysing rainfall data collected in the form of a tipping-bucket time series.

In this paper, we describe a new point process model which uses exponentially decaying rainfall pulses associated with each event of the rainfall burst resulting from a doubly stochastic Poisson process. We study some of the second-order properties of the rainfall aggregations in discrete time intervals and use the method of moment approach to estimate parameters of the proposed model. Empirical and fitted values of the properties are compared. Simulation from the fitted model has been used in model assessment.

Ramesh, N.I., Onof, C. and Xie, D. (2012). Doubly Stochastic Poisson Process models for precipitation at fine time-scales, Advances in Water Resources, 2012; 45: 58-64. Ramesh, N. I, Garthwaite, A and Onof, C (2017). A doubly stochastic rainfall model with exponentially decaying pulses. Stochastic Environmental Research and Risk Assessment, 32 (6). pp. 1645-1664.

Data Analysis in the Environment of Destructive Samples: The Case of Spina Bifida

Tuesday 23
10:30 - 12:30
Room 8

Marepalli Rao[1] and Tianyuan Guan

[1]Biostatistics and Bioinformatics, Cincinnati, United States
marepalli.rao@uc.edu

Spina Bifida is a birth defect. At fetus level, skin on the spinal column is not fully formed. Spinal fluid is seeping out of the spinal column through the gaps on the column. Amniotic fluid is getting into the spinal column. This will create many medical problems if the fetus survives. Fetoscopy is an option. A biomedical engineering team developed a biodegradable patch made out of polymers to cover the gaps. Hopefully, skin forms under the patch. It is important to measure how rough the patch is over time. The rougher the patch is, the better it holds nutrients helpful for formation of skin. The process that is used to measure roughness destroys the patch. A critical statistical problem that arose in this context is to predict what roughness is at 0 weeks for the patch that is removed at 4 weeks, for example. We have proposed a novel bootstrap procedure to solve the problem.

Circuits in Experimental Design

Tuesday 23
14:00 - 16:00
Room 5

Fabio Rapallo[1]

[1]Department DISIT, University of Piemonte Orientale, Alessandria, Italy
fabio.rapallo@uniupo.it

In the framework of factorial design, a statistical (linear) model is defined through an appropriate model matrix, which depends on the experimental runs and encodes their geometric structure. In this seminar we descuss some properties of the circuit basis of the model matrix in connection with two well-known properties of the designs, namely robustness and D-optimality. Exploiting the identification of a fraction with a binary contingency table, we define a criterion to check whether a fraction is saturated or not with respect to a given model and we generalize such a result in order to study the robustness of a fraction by inspecting its intersections with the supports of the circuits. Using some simulations, we show that the combinatorial description of a fraction with respect to the circuit basis is strictly related to the notion of D-optimal fraction and to other optimality criteria. This talk is based on joint work with Henry Wynn.

Tuesday 23
16:30 - 17:30
Second Floor

Causal Quantile Learner: Causal Inference for Structural Equation Model

Denise Rava[1] and Jelena Bradic

[1]Department of Mathematics, University of California, San Diego, United States
drava@ucsd.edu

Discovering cause-effect relationships between variables is a fundamental but very challenging task that finds its application in various area of interest. We focus here on the setting where both observational and interventional data are available. The latter are generated by different type of interferences. In this contest, a possibility is to use Invariant Causal Prediction methodology or Instrumental Variable type of methods. In the spirit of the former, we propose here two robust methods for estimation of the causal structure, one suitable for low dimensional data, Causal Quantile Learner and one for high dimensional one, Causal Quantile Dantzig. Both are tailored for the setting where data from multiple environments are available and the main idea is to use a notion of invariance between different environments to estimate the underlying causal structure. We introduce a new type of invariance, called Quantile Invariance, that makes use of the quantile loss assuming that the dependency structure of the quantile of the target variable Y on the predictors X remains stable between environments. We show that, in practice, this invariance holds in the context of linear SEM with possible hidden confounders, under both additive and multiplicative interventions. We construct two estimators that inherit the desirable robustness of quantile regression and that can be used in this more complicated framework of multiple environments to answer causal inference type of questions. Our simulations show that Causal Quantile Learner and Causal Quantile Dantzig outperforms already proposed methodologies in terms of predictive power and estimation error under a broad spectrum of heavy-tailed data generating processes.

Friday 26
09:00 - 11:00
Room 9

Nonparametric Statistical Inference for the Drift of a Multidimensional Diffusion

Kolyan Ray[1] and Richard Nickl

[1]King's College London, London, United Kingdom
kolyan.ray@kcl.ac.uk

We consider the problem of estimating the drift and invariant measure of a periodic multidimensional diffusion based on continuous observations, a model used for instance in molecular dynamics. Placing a high dimensional Gaussian prior on the drift, we obtain convergence rates for the Bayesian posterior distribution and the corresponding maximum a posteriori (MAP) estimate, which equals a penalized least squares estimator.

For dimension at most 3, we further obtain Bernstein-von Mises type results for the posterior distributions of both the drift and invariant measure. This provides a frequentist justification for the Bayesian approach in this model, including for uncertainty quantification.

Multi-Output Chain Models and their Application in Data Streams

Thursday 25
14:00 - 16:00
Room 5

Jesse Read[1]

[1]Ecole Polytechnique, Palaiseau, France
jesse.read@polytechnique.edu

The method of Classifier Chains uses predictions of models as additional features to the following models across a 'chain' of related target variables. It has proven successful in multi-label classification tasks as well as many related fields. When employed to predict real-valued targets, the analogue (Regressor Chains) behave differently, depending on the chosen loss function and base models. We explore the development of probabilistic multi-output chain methods, in particular drawing inspiration from Particle Filters. We also study the application of such methods to tasks such as missing value imputation and anomaly detection in dynamic data streams, which show promising results.

Properties of the Stochastic Approximation EM Algorithm with Mini-Batch Sampling

Tuesday 23
10:30 - 12:30
Room 6

Tabea Rebafka[1] and Estelle Kuhn and Catherine Matias

[1]LPSM, Sorbonne University, Paris, France
tabea.rebafka@upmc.fr

For models where the classical EM algorithm cannot be applied directly, stochastic variants such as Monte Carlo EM, Stochastic Approximation EM (SAEM) and Monte Carlo Markov Chain SAEM (MCMC-SAEM) exist. However, their computing time is very long when the sample size and hence the number of latent variables is large. As a solution mini-batch sampling has been proposed recently, which consists in using only a part of the observations and simulating only a portion of the latent variables at each iteration. Intuitively, when the so-called mini-batch size, that is the size of the data subset selected at every iteration, is small, the computing time is shortened, while the computed estimator may be less accurate.

In this talk, we propose a mini-batch version of the MCMC-SAEM algorithm, which is appropriate when the latent data cannot be simulated exactly from the conditional distribution, as for instance in nonlinear models or non-Gaussian models. As the underlying stochastic approximation procedure only requires the simulation of a single instance of the latent variable at every iteration, MCMC-SAEM is much more computing efficient than MCMC-EM. Nevertheless, when the dimension of the latent variables is huge, the sampling step can still be time-consuming and thus our mini-batch version is computationally more efficient than the original algorithm.

When the model belongs to the exponential family, we prove almost-sure convergence of the sequence of estimates generated by the mini-batch MCMC-SAEM algorithm as the number of iterations increases. Moreover, we provide results in the same regime that quantify the impact of the mini-batch size on the limit distribution of the estimator compared to the classical batch MCMC-SAEM algorithm. Simulation experiments and real data examples show that an appropriate choice of the mini-batch size results in an important speed-up of

the convergence in nonlinear mixed effects models, frailty models and the stochastic block model.

Monday 22
16:30 - 18:30
Room 8

Autoregressive-Type Time Series Models with Bounded Support

Lena Reichmann[1] and Carsten Jentsch

[1]Mannheim, Germany
reichmann@uni-mannheim.de

In practice, many time series are bounded from below and above such that they can take only values in a certain interval [a,b] instead of the whole real line. Such cases include rates, weights and proportions. Most modelling approaches for such time series rely on certain transformations that map the interval [a,b] to the real line such that classical time series models, as e.g. autoregressive moving-average models, can be applied. However, such a transformation is clearly not uniquely determined such that the model will generally depend on the specific transformation that has been used. By considering a direct modeling approach that avoid transformations to capture the serial dependence, we propose a new autoregressive moving average type model class with nicely interpretable structure. To take the boundedness of such time series into account, the proposed models make use of beta distributions. In Rocha and Cribary – Neto (2008), the authors followed a similar path and proposed the beta ARMA model class, which also relies on beta distributions conditional on the past, but have to utilize also a suitable link function to transform the response to the unit interval. For our new model class, we provide sufficient stationarity conditions and derive the stationary solution of the model equations. For the purely autoregressive case, we prove the Yule-Walker equations to hold which facilitate the task of parameter estimation in these new models as the whole toolbox for classical autoregressive models becomes applicable. Further, we discuss mixing properties and provide some simulation results.

Monday 22
14:00 - 16:00
Room 12

Minimax Rates of Estimation for Smooth Optimal Transport Maps

Jan-Christian Hütter and Philippe Rigollet[1]

[1]Department of Mathematics, Massachusetts Institute of Technology, USA
rigollet@math.mit.edu

Brenier's theorem is a cornerstone of optimal transport that guarantees the existence of an optimal transport map T between two probability distributions P and Q over $\mathbb{R}^d$ under certain regularity conditions. In this presentation we show how to establish the first minimax rates estimation rates for such a transport map from data sampled from P and Q under additional smoothness assumptions on T.

Nonparametric Estimation for Size-Structured Population of Cells

Tuesday 23
14:00 - 16:00
Room 9

Vincent Rivoirard[1]

[1]Paris-Dauphine University, Paris, France
Vincent.Rivoirard@dauphine.fr

We consider a stochastic individual-based model in continuous time to describe a size-structured population for cell divisions. This model is motivated by the detection of cellular aging in biology. The size of the system evolves according to a transport-fragmentation equation: each individual grows with a given transport rate and splits into two offsprings. In the first part of this talk, we assume that offsprings have the same size, following a binary fragmentation process with unknown division rate that depends on its size. Our nonparametric procedure to deal with the problem of rate estimation relies on kernel methods with automatic bandwidth selection performed by the Goldenshluger-Lepski methodology. In the second part, we do no longer assume that offsprings have the same size and we then address the problem of nonparametric estimation of the kernel ruling the divisions based on the eigenvalue problem related to the asymptotic behavior in large population. This inverse problem involves a multiplicative deconvolution operator. Using Fourier technics we derive a nonparametric estimator whose consistency is studied.

Maximum Likelihood Estimation for Totally Positive Densities

Tuesday 23
14:00 - 16:00
Room 5

Elina Robeva[1] and Bernd Sturmfels and Ngoc May Tran and Caroline Uhler

[1]Mathematics, Massachusetts Institute of Technology, United States
erobeva@mit.edu

Nonparametric density estimation is a challenging problem in theoretical statistics – in general the maximum likelihood estimate (MLE) does not even exist! Introducing shape constraints allows a path forward. This talk offers an invitation to non-parametric density estimation under total positivity (i.e. log-supermodularity) and log-concavity. Totally positive random variables are ubiquitous in real world data and possess appealing mathematical properties. Given i.i.d. samples from such a distribution, we prove that the maximum likelihood estimator under these shape constraints exists with probability one. We characterize the domain of the MLE and show that it is in general larger than the convex hull of the observations. If the observations are 2-dimensional or binary, we show that the logarithm of the MLE is a tent function (i.e. a piecewise linear function) with "poles" at the observations, and we show that a certain convex program can find it. In the general case the MLE is more complicated. We give necessary and sufficient conditions for a tent function to be concave and supermodular, which characterizes all the possible candidates for the MLE in the general case.

Wednesday 24
10:30 - 12:30
Room 6

The Influence of Sponsorship in the Intention Purchase of Portuguese Consumers

Helena Sofia Rodrigues[1] and Manuel J. S. Pereira and António Macedo

[1]Business School, Polytechnic Institute of Viana do Castelo, Valença, Portugal
sofiarodrigues@esce.ipvc.pt

In a competitive business environment, companies are increasing their quest for efficiency, demanding more marketing benefits and, at the same time tangible and measurable results that justify their investments. Sponsorship, as a marketing tool explored in business context, could be a channel that enables the achievement of strategic objectives that are advantageous to the organization, in terms of institutional, communicational and commercial aspects. This research sought to understand how sponsorship could impact the purchase intent of consumers about the products and services of a sponsoring brand. It is also intended to explore the effects of the relationship about the sponsoring entity with its sponsorship activities and with the sportive events in which the brand is involved. In this work are also explored concepts linked to the consumer behavior with the sponsor, such as the involvement, the attitude and the purchase intent. Therefore, a research methodology focused on the questionnaire survey was carried out. The results of this investigation allowed to verify that the relationship between the Oakley brand with its sponsorship activities and also with the sports events in which the brand is involved, reflects an advantageous and profitable relationship, intend by the consumer. Positive results related to consumer involvement and attitude toward Oakley brand were also found. Regarding the dimension based on purchase intention of the consumer about the products and services of Oakley brand, there was identified a constructive relationship and positive influence of the dimensions mentioned above, in this specific component. This feature increases the chance of the consumer to test and to buy the brand products and services.

Tuesday 23
16:30 - 17:30
Second Floor

An Economic Application of the CTP Distribution

Jose Rodríguez-Avi[1] and Julia Rodríguez-Reinoso and Valentina Cueva-López and María José Olmo-Jiménez

[1]Estadística e Investigación Operativa, Jaén, Spain
jravi@ujaen.es

The complex triparametric Pearson (CTP) distribution is a count data distribution that can be used to model overdispersed as well as underdispersed discrete data sets. In the case of overdispersion (the variance greater than the mean) many models have been proposed, such as Negative binomial, Generalized Poisson or Waring, among other, but these models assume that this overdispersion is due to an higher number of zeros than the proposed by the Poisson distribution with the same mean. Nevertheless, the CTP distribution allows us to model overdipersed data when this characteristic is due to a high presence of low but not zero values in the distribution.

These type of data have an economical reflect if we describe the variable number of facilities (in any form) by municipality. If we analyse the profile of these variables, we can observe a difference in respect with the type of ownership, public or private. In the private case

the number of facilities are related to the economic efficiency. In consequence, the number of 0 facilities will be high and the variable can be modelled through the use of discrete distributions obtained as a Poisson mixture.

Nevertheless, for the public ownership case the number of facilities is not related only with economic, but other criteria. In these cases, the overdispersion is due to a greater probability of ones (at least one school, library, walk-in clinic. . .) , whereas the probability of zero is nearer, or even lower, than to the probability given by a Poisson with the same mean. In this case, the CTP distribution is the more adequate to propose a statistical model for such class of data.

In this work we show how this distribution works in several examples where different public facilities are fitted by the CTP and the remaining distributions. We analyze the shape of these models and show how the CTP gives the best fits according to several criteria. Data have been obtained from different Autonomic communities in Spain.

Statistical Space-Time Projections of Wave Heights in the North Atlantic

Monday 22
16:30 - 18:30
Room 10

Hanne Rognebakke[1] and Thordis Thorarinsdottir and Hugo Hammer and Tor Arne Øigård

[1]Norwegian Computing Center, Oslo, Norway
hanne.rognebakke@nr.no

Future changes in ocean wind wave climate have broad implications for the physical infrastructure and environment in coastal, near- and offshore regions. A comprehensive assessment of the impact of climate change and the associated risks in these regions thus requires the consideration of changes in wave climate. However, the understanding of projected changes in wave climate is limited relative to many other climate variables such as temperature and precipitation. This results from the fact that coupled atmosphere-ocean general circulation models (GCMs) typically lack wind wave parameterizations so that wave parameters are not available as a standard model output. To decrease this knowledge gap, we have developed new statistical models for wave climate as well as tools for the analysis of wave climate projections. We demonstrate that non-stationary, fuzzy time series approaches can be applied to generate projections of future wind and waves. Further, by using regional extreme value analysis, we show that regional estimates of extreme wave height quantiles are more robust and associated with a smaller uncertainty than those commonly obtained from local analyses. Generally, future wave climate projections are associated with a large uncertainty and an overall decreasing trend in the Northern Atlantic is observed.

Monday 22
14:00 - 16:00
Room 9

Is Climate Change Making Extreme Rains more Frequent, or Bigger, or more Dangerous?

Holger Rootzén[1] and Helga Olafsdottir and David Bolin

[1]Mathematical Sciences, Gothenburg, Sweden
hrootzen@chalmers.se

Records can become bigger because the underlying distribution changes, or because one makes more tries, or both. Dangers of flooding is determined by the interplay between changes in record sizes and the spatial extents of extreme rainfall events. We use NOAA data with more than 60 years of observation, and with observations going up to at least 2010, to study large rainstorms in northeastern USA. Data on individual rainstorms provide the most direct path to understanding the development of rainstorms. However, annual maxima data are more widely available, and often of higher quality. We hence use both kinds of data, together with the close relation between the PoT method with generalized Pareto distributed excesses and the annual maxima method with the generalized extreme value distribution. This relation is closely related to Langbein's formula, which is widely used in hydrology to connect partial duration series with annual maxima. A preliminary answer is that rainstorms in northeastern USA are becoming more frequent, but that the distribution of the amount of rain in individual rainstorm is not changing. To model the spatial extent and changes in spatial extent of rainfall events we derive new explicit representations of the GP distributions associated with the stable mixture GEV distributions. These models inherit the appealing properties of the stable mixture GEV models which provide components of variance models, time series models, and spatial and continuous parameter models for annual maxima. Additionally, they allow for simple explicit formulas for densities, and hence for easier ML estimation than competing models. An important challenge is that extremes, say rainstorms, may hit a some, but not all, spatial locations, so that some, but not all, components are extreme. For GP modelling this means that models must be able to accommodate cases where some of the components have mass on their lower boundary, which may be finite or equal to minus infinity.

Monday 22
09:00 - 10:00
Aula Magna
Building 13

Bayesian Measures of Uncertainty

Judith Rousseau[1]

[1]University of Oxford, Oxford, United Kingdom
udith.rousseau@stats.ox.ac.uk

The reknown theorem of Bernstein von Mises in regular finite dimensional models has numerous interesting consequences, in particular it implies that a large class of credible regions are also asymptotically confidence regions, which in turns imply that different priors lead to the same credible regions to first order.

Unfortunately the Bernstein von Mises theorem does not necessarily hold in high or infinite dimensional models and understanding the asymptotic behaviour of credible regions is much more involved. In this talk I will describe what are the new advances that have been obtained over the last 8 years or so on the understanding - or not- of credible regions in semi and non- parametric models.

I will in particular discuss some interesting phenomena which have been exhibited in high dimensional models, for certain families of priors, encountered for instance in mixture models with unknown number of components, in regression models with a large number of covariates etc... We can show that in a significant number of cases these priors tend to over penalize (or over smooth), leading to only partially robust confidence statements.

I will also discuss the few advances which have been obtained in the context of non or semi parametric mixture models, which are notoriously difficult to study.

Strong Approximations of Brownian Sheet by Uniform Transport Processe

Tuesday 23
16:30 - 17:30
Second Floor

Carles Rovira[1] and Xavier Bardina and Marco Ferrante

[1]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
carles.rovira@ub.edu

Many years ago, Griego, Heath and Ruiz-Moncayo proved that it is possible to define realizations of a sequence of uniform transform processes that converges almost surely to the standard Brownian motion, uniformly on the unit time interval. In this paper we extend their results to the multi parameter case. We begin constructing a family of processes, starting from a set of independent standard Poisson processes, that has realizations that converge almost surely to the Brownian sheet, uniformly on the unit square. At the end the extension to the d-parameter Wiener processes is presented.

Sparsely Observed Functional Time Series: Estimation and Prediction

Monday 22
10:30 - 12:30
Room 10

Tomas Rubin[1] and Victor M. Panaretos

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Czech Republic
tomas.rubin@epfl.ch

Functional time series analysis has traditionally been carried out under the assumption of complete observation of the constituent series of curves, assumed stationary. Nevertheless, it may well happen that the data available to the analyst are not the actual sequence of curves, but relatively few and noisy measurements per curve, potentially at different locations in each curve's domain. The subject is to tackle the problem of estimating the dynamics and of recovering the latent process of smooth curves in this sparse observation regime. We construct a consistent nonparametric estimator of the series' spectral density operator and use it develop a frequency-domain recovery approach, that predicts the latent curves at a given time by borrowing strength from the (estimated) dynamic correlations in the series across time. Further to predicting the latent curves from their noisy point samples, the method fills in gaps in the sequence (curves nowhere sampled), denoises the data, and serves as a basis for forecasting. Means of providing corresponding confidence bands are also investigated. The methodology is further illustrated by application to an

environmental data set on fair-weather atmospheric electricity, which naturally leads to a sparse functional time-series.

Tuesday 23
14:00 - 16:00
Room 11

Scaling Up Optimal Kernel Methods for Large Scale Machine Learning

Alessandro Rudi[1]

[1]INRIA, Sierra team and École Normale Supérieure, Paris
Alessandro.Rudi@ens.fr

Kernel methods provide a principled way to perform non linear, nonparametric learning. They rely on solid functional analytic foundations and enjoy optimal statistical properties. However, at least in their basic form, they have limited applicability in large scale scenarios because of stringent computational requirements in terms of time and especially memory. In this paper, we take a substantial step in scaling up kernel methods, proposing FALKON, a novel algorithm that allows to efficiently process millions of points. FALKON is derived combining several algorithmic principles, namely stochastic subsampling, iterative solvers and preconditioning. Our theoretical analysis shows that optimal statistical accuracy is achieved requiring essentially O(n) memory and O(n sqrt(n)) time. An extensive experimental analysis on large scale datasets shows that, even with a single machine, FALKON outperforms previous state of the art solutions, which exploit parallel/distributed architectures.

Thursday 25
14:00 - 16:00
Room 5

A Contrastive Divergence for Combining Variational Inference and MCMC

Francisco Ruiz[1] and Michalis Titsias

[1]University of Cambridge / Columbia University, Cambridge, United Kingdom
f.ruiz@columbia.edu

We develop a method to combine Markov chain Monte Carlo (MCMC) and variational inference (VI), leveraging the advantages of both inference approaches. Specifically, we improve the variational distribution by running a few MCMC steps. To make inference tractable, we introduce the variational contrastive divergence (VCD), a new divergence that replaces the standard Kullback-Leibler (KL) divergence used in VI. The VCD captures a notion of discrepancy between the initial variational distribution and its improved version (obtained after running the MCMC steps), and it converges asymptotically to the symmetrized KL divergence between the variational distribution and the posterior of interest. The VCD objective can be optimized efficiently with respect to the variational parameters via stochastic optimization. We show experimentally that optimizing the VCD leads to better predictive performance on two latent variable models: logistic matrix factorization and variational autoencoders (VAEs).

Data-Driven Support Estimation

Tuesday 23
14:00 - 16:00
Room 12

Paula Saavedra-Nieves[1] and Alberto Rodríguez-Casal

[1]Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Santiago de Compostela, Spain
paula.saavedra@usc.es

This work deals with the problem of estimating the compact support S of an absolutely continuous random vector X from an independent and identically distributed sample of observations taken in it.

Several proposals for reconstructing S have been considered in the literature. For instance, Devroye-Wise (1980) proposed to estimate S as a union of closed balls with centers in sample points and radius depending only on the sample size. More sophisticated estimators can be used if we have some additional information on the set. For example, if we know that S is convex then the convex hull of the sample is a natural support estimator. But convexity assumption may be too restrictive. In this case, it is necessary to consider more flexible shape restrictions such as r-convexity for $r>0$. If we assume that S is r-convex then the r-convex hull of the sample provides a reasonable reconstruction of the support. In Rodríguez-Casal (2007), it is proved that if the smoothing parameter r is correctly chosen, the r-convex hull of the sample achieves the same convergence rates (in Hausdorff and distance in measure) as the convex hull. But, in practice, S is unknown and, consequently, the real value of the geometric index r too. Rodríguez-Casal and Saavedra-Nieves (2016) proposed a stochastic algorithm for selecting it from the data under the hypothesis that the sample is uniformly generated. The resulting reconstruction of S is able to achieve the same convergence rates as the convex hull for estimating convex sets.

In this work, we will extend the previous result. We propose a consistent estimator of the largest value of r such that S is r-convex under the assumption that the sample is generated from a density that is bounded from below and Lipschitz continuous restricted to its bounded support S. The resulting estimator of S also achieves the same convergence rates as the convex hull for estimating convex sets, but under much more flexible sampling scenario.

References

Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions, l'I.H.P.-Probabilites & Statistiques, 43, 763-774.

Rodríguez-Casal, A. and Saavedra-Nieves, P. (2016). A fully data-driven method for estimating the shape of a point cloud, ESAIM: Probability and Statistics, 20, 332-348.

Tuesday 23
14:00 - 16:00
Room 10

Misclassification-Robust Semiparametric Estimation of Single-Index Binary Choice Models

Serhan Sadikoglu[1] and Pavel Cizek

[1]Department of Econometrics and Operations Research, Tilburg University, Tilburg, Netherlands
s.sadikoglu@uvt.nl

In this paper, a new class of semiparametric estimators for single-index binary-choice models is introduced. The proposed estimators are based on the semiparametric indirect inference that suggests estimating the parameters of the model via possibly misspecified auxiliary criteria. A large class of considered auxiliary criteria includes the ordinary least squares, nonlinear least squares, and nonlinear least absolute deviations estimators. Besides deriving the consistency and asymptotic normality of the proposed methods, we demonstrate that the proposed indirect inference methodology - at least for selected auxiliary criteria-combines weak distributional assumptions, good estimation precision, and robustness to misclassification of responses. We conduct Monte Carlo experiments to compare the finite-sample performance of the proposed and existing estimators.

Tuesday 23
16:30 - 17:30
Second Floor

ayesian Test of Bimodality for the Generalized von Mises Distribution

Sara Salvador[1] and Riccardo Gatto

[1]Institut für Mathematische Statistik und Versicherungslehre, University of Bern, Bern, Switzerland
sara.salvador@stat.unibe.ch

The generalized von Mises (GvM) is a distribution on the circle and it involves circular data that can represents, for example, wind directions or vanishing angles at the horizon for a group of birds. The importance of knowing the number of modes for a circular distribution is straightforward: for example in the case of circular data involving directions of pigeons, unimodality would imply that the birds have a preferred vanishing direction, and this could be of considerable scientific interest. The aim of the project is to study the bimodality of the GvM by computing the Bayesian Factor in order to make a decision between the hypothesis “the distribution is bimodal” against “the distribution is unimodal”. The approach is Bayesian: we choose as priors for the concentration parameters κ_1 and κ_2 two uniforms distributions, while for the mean directions μ_1 and μ_2 we choose von Mises and bimodal von Mises distributions respectively. As likelihood function we consider a GvM and thus, using the Bayes theorem we obtain the posterior distribution. The choice between the two hypothesis is computed comparing the prior and posterior probability of bimodality. For the prior case, we simulate component-wise from the prior distributions of the parameters and count the cases that gives bimodality. For the posterior case, the procedure is similar but the simulation from the posterior distribution is hard and then we use a Gibbs sampler that simulates from the posterior without directly using it: the simulation is computed instead from conditional distributions, called full conditionals, using a Metropolis-Hastings algorithm. The conclusions confirm the tendency for the Generalized von Mises of being bimodal.

Nonparametric Independence Testing via Mutual Information

Monday 22
16:30 - 18:30
Room 9

Richard Samworth[1] and Thomas Berrett

[1]Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom
r.samworth@statslab.cam.ac.uk

We propose a test of independence of two multivariate random vectors, given a sample from the underlying population. Our approach is based on the estimation of mutual information, whose decomposition into joint and marginal entropies facilitates the use of recently-developed efficient entropy estimators derived from nearest neighbour distances. Our critical values, which may be obtained from simulation (in the case where one marginal is known) or resampling, guarantee that the test has nominal size, and we provide a local power analysis, uniformly over classes of densities whose mutual information satisfies a lower bound. Our ideas may be extended to provide a new goodness-of-fit tests of normal linear models based on assessing the independence of our vector of covariates and an appropriately-defined notion of an error vector.

A Bayesian Non-Homogeneous Markov Chains to Modeling and Analyzing Multiple Sclerosis Progression

Tuesday 23
16:30 - 17:30
Second Floor

Istoni Sant'Ana[1] and Astrid Quinones and Manueli Centeno and Angel Chinea-Martinez

[1]Biostatistics and Epidemiology, University of Puerto Rico, San Juan, United States
istoni.daluz@upr.edu

Multiple sclerosis (MS) is a chronic neurological autoimmune disease that most commonly affects young adults among ages of 20 and 40 years old. MS are considered as one of the most prevalent neurological disorders and causes of disability among the young adult population worldwide. The purpose of this study is to model the progression of MS and the effect of its therapies measured with disability measure scale. It is a typical event whose progression is non-homogeneous. For this reason, we suggest a non-homogeneous Markovian structure to describe the transition patters for patients, incorporating time-dependent covariates and controlling by confounding variables. In modelling this type of data, a parameter-driven hierarchical Bayesian framework was proposed. Inference approaches are addressed to issues found in the analyses of data from multiple sclerosis treatment programs. The model and inference procedures are used in real life longitudinal data, reported by San Juan MS Center patients.

Tuesday 23
16:30 - 17:30
Second Floor

Robust Multivariate Estimation Based on Statistical Depth Filters

Giovanni Saraceno[1] and Claudio Agostinelli

[1]Department of Mathematics, Università degli studi di Trento, Trento, Italy
giovanni.saraceno@unitn.it

In the classical contamination models, such as the gross-error (Huber and Tukey contamination model or Case-wise Contamination) observations are considered as the units to be identified as outliers, this model is very useful when the number of considered variables is moderately small. Alquallaf et al. [2009] shows the limits of this approach for a larger number of variables and introduced the Independent Contamination model (Cell-wise Contamination) where now the cells are the units to be identified as outliers or not. One approach to deal at the same time with both types of contamination is filter out the outliers from the data set and then apply a robust procedure able to handle missing values. Here we develop a general framework to build filters in any dimension based on statistical data depth functions. We show that previous approaches such as the one presented in, e.g., Leung et al. [2017] are special cases. We illustrate the procedure by using the halfspace depth function. This is basically a two-step procedure: first, apply the depth filter to the data matrix and set the flagged cells to missing values, NA's, and second, apply the generalized S-estimator (GSE) of Danilov et al. [2012] to the incomplete dataset. We illustrate the features of our approach using a real data set and we performed a Monte Carlo simulation to assess the performance of our procedure. References F. Alqallaf, S. Van Aelst, R. H. Zamar, and V. J. Yohai. Propagation of outliers in multivariate data. The Annals of Statistics, 37(1):311–331, 2009. M. Danilov, V.J. Yohai, and R.H. Zamar. Robust estimation of multivariate location and scatter in the presence of missing data. Journal of the American Statistical Association, 107:1178–1186, 2012. A. Leung, V.J. Yohai, and R.H. Zamar. Multivariate location and scatter matrix estimation under cellwise and casewise contamination. Computational Statistics and Data Analysis, 111:59–76, 2017.

Tuesday 23
14:00 - 16:00
Room 10

Robust to Outliers of Median-of-Means

Adrien Saumard[1] and Edouard Genetay

[1]CREST, ENSAI, Université Bretagne Loire, Bruz, France
asaumard@gmail.com

Nowadays, datasets are often massive and messy, containing errors of measurements, missing data or even being voluntary corrupted by a cyber-attack. This naturally led to an intense activity in statistical learning and (high-dimensional) statistics, revisiting and adapting robustness issues to modern challenges. A systematic approach to robustness in machine learning has developed through the use along the algorithms of the median-of-means statistics, that aims at providing a robust mean estimation by a genuine divide-and-conquer strategy. We will give a sharp analysis of the behavior of the meadian-of-means - and some of its variants - with respect to outliers, starting from the computation of its breakdown point and then introducing some more refined concepts of robustness.

A New Multivariate Two-Sample Rank Test

Tuesday 23
16:30 - 17:30
Second Floor

Martin Schindler[1]

[1]Department of Applied Mathematics, Technical University of Liberec, Liberec, Czech Republic
martin.schindler@tul.cz

A two sample multivariate testing problem is considered. Jureckova and Kalina (2012) proposed several rank tests which are unbiased against a broad class of alternatives and are distribution free. Among others they illustrated the good properties of their Wilcoxon-type test on simulations compared to other test. Here a test statistic which is a modification of this Wilcoxon-type test is considered. The test statistic is based on the ranks of inter-point distances of the multivariate observations. Based on the proposed test statistic a permutation test is introduced. Its performance is illustrated on simulated data of different distribution. The powers of the proposed test is compared to the Wilcoxon-type test and to a classical test. The proposed test mostly dominates the other two regardless of the distribution of the data.

Advanced Models and Methods for Bivariate Random Fields

Monday 22
16:30 - 18:30
Room 11

Martin Schlather[1] and Olga Moreva and Niklas Hansen

[1]Institute of Mathematics, Universität Mannheim, Mannheim, Germany
schlather@math.uni-mannheim.de

Although the area of geostatistics is more than fifty years old, still a sufficiently broad spectrum of covariance models is not known yet. For instance, in the multivariate case, particularly bivariate case, only a few classes of models have been found. The talk gives an overview over the currently known models and its simulation methods. Some model approaches are extended, such as Gneiting's fully symmetric bivariate Whittle-Matern model and Wackernagel's delay model to get new classes of models. Advanced simulation techniques, such as the cutoff circulant embedding, are extended to the bivariate case. Finally, the statistical inference with the R package RandomFields is presented.

Friday 26
09:00 - 11:00
Room 5

Latent Class Models for Diagnostic Testing with No Gold Standard

Matthew Schofield[1] and Katrina Sharples and Michael Maze and John Crump

[1]University of Otago, Dunedin, New Zealand
matthew.schofield@otago.ac.nz

We explore the use of latent class models for diagnostic testing when there is no gold standard. The work is motivated by a study of Leptospirosis in Tanzania, where the four possible testing procedures were considered. A simple two-state latent class model fitted well, but the sensitivities obtained were counter-intuitive; the test expected to perform best, had the lowest sensitivity. We show using simulation that the assumption that the latent class corresponds to disease status can be problematic. This can lead to large bias in the estimated sensitivities while having minimal effect on the fit of the model.

Thursday 25
10:30 - 12:30
Seminari B

Convergence Rates for the (Generalized) Fréchet Mean via the Quadruple Inequality

Christof Schötz[1]

[1]Heidelberg University, Heidelberg, Germany
schoetz@math.uni-heidelberg.de

The Fréchet mean generalizes the notion of mean to random objects in arbitrary metric spaces. We provide convergence rates in probability and in expectation for the empirical Fréchet mean. In contrast to previous results on Fréchet means, we do not require a finite diameter of the metric space. Instead, we assume an inequality, which we call quadruple inequality, and by that, generalize previous results. The quadruple inequality is known to hold in Hadamard spaces of any diameter. We show that it also holds in a suitable way for certain powers of a Hadamard-metric.

Monday 22
10:30 - 12:30
Room 5

Normal Approximation of Stabilising Functionals

Matthias Schulte[1]

[1]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland
matthias.schulte@stat.unibe.ch

Many statistics of point processes can be written as sums of scores, where each score represents the contribution of a point. If such a score depends only on the point process in a random neighbourhood of the point, the statistic is called stabilising. For stabilising functionals of underlying Poisson or binomial point processes central limit theorems with presumably optimal rates of convergence are presented. Statistical applications such as Voronoi set approximation are considered.

A Weighted Fshape Model in Computational Anatomy

Tuesday 23
10:30 - 12:30
Room 7

Tomonari Sei[1] and Tomoki Tanaka and Shouhei Hanaoka and Hidekata Hontani

[1]The University of Tokyo, Tokyo, Japan
sei@mist.i.u-tokyo.ac.jp

The functional-shape (fshape) model is a statistical model for registration between two manifolds equipped with signal functions (Charlier, Charon and Trouve, 2017, Foundations of Computational Mathematics, 17 (2), 287–357). We present an extension of the fshape model, where a weight function of environment is introduced to take into account the anatomical prior knowledge. Numerical experiments show that variability of signals is controlled by the weight function as expected. This research is a joint work with Akinobu Shimizu.

RKHS Mean Embeddings and Hypothesis Testing

Friday 26
09:00 - 11:00
Room 10

Dino Sejdinovic[1]

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom
dino.sejdinovic@stats.ox.ac.uk

Kernel embeddings of distributions into reproducing kernel Hilbert spaces and the Maximum Mean Discrepancy (MMD), the resulting distance between distributions, are useful tools for fully nonparametric hypothesis testing and for learning on distributional inputs. I will give an overview of this framework and present some of its recent applications and extensions.

Manifold Valued Data Analysis of Samples of Networks, with Applications in Corpus Linguistics

Tuesday 23
10:30 - 12:30
Room 7

Katie Severn[1] and Ian Dryden and Simon Preston

[1]Mathematics, University of Nottingham, Nottingham, United Kingdom
pmxks10@nottingham.ac.uk

Networks can be used to represent many systems such as text documents and brain activity, and it is of interest to develop statistical techniques to compare networks. We develop a general framework for extrinsic statistical analysis of samples of networks, motivated by networks representing text documents in corpus linguistics. We identify networks with their graph Laplacian matrices, for which we define metrics, embeddings, tangent spaces, and a projection from Euclidean space to the space of graph Laplacians. This framework provides a way of computing means, performing principal component analysis and regression, and performing hypothesis tests, such as for testing for equality of means between two samples of networks. We apply the methodology to the set of novels by Jane Austen and Charles Dickens.

Tuesday 23
16:30 - 17:30
Second Floor

Bayesian Non-Parametric Inference for Stochastic Epidemic Models

Rowland Seymour[1] and Theodore Kypraios and Philip O'neill

[1]University of Nottingham, Nottingham, United Kingdom
rowland.seymour@nottingham.ac.uk

Simulating from and making inference for stochastic epidemic models are key strategies for understanding and controlling the spread of infectious diseases. Despite the enormous attention given to methods for parameter estimation, there has been relatively little activity in the area of non-parametric inference. That is, drawing inference for the infection rate without making specific modelling assumptions about its functional form. We develop novel Bayesian non-parametric methodology to fit heterogeneously mixing models in which the infection rate between two individuals is a function, $f(\cdot)$, of their characteristics, for example location or type. Making non-parametric inference in this context is very challenging because the likelihood function of the observed data is intractable. We adopt a fully Bayesian approach by assigning a Gaussian Process (GP) prior to $f(\cdot)$ and then develop an efficient data augmentation Markov Chain Monte Carlo methodology to estimate $f(\cdot)$, the GP hyperparameters and the unobserved infection times. We then extend this method by using multi-output GP prior distributions to infer infection rates which depend on both continuous and discrete characteristics and covariates. We illustrate our methodology using simulated data and by analysing a data set on Avian Influenza from the Netherlands.

Monday 22
10:30 - 12:30
Room 5

Recent Progress of Self-Normalized Limit Theory

Qi-Man Shao[1]

[1]Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China
shaoqm@sustech.edu.cn

The standardized coefficients in the classical limit theorems are usually deterministic , while they are random in the self-normalized limit theorems. Typical examples of self-normalization include the well-known student t-statistic and studentized U-statistics and studentized non-linear statistics. It has been proved that the tail probability of studentized statistics is often more robust than that of the standardized statistics. For example, the large deviation for self-normalized sums of independent and identically distributed random variables holds without any moment assumptions and the self-normalized Cramer type moderate deviations holds under a finite third moments. In this talk we will give a brief survey on recent progress of self-normalized limit theory.

Hug and Hop: Explicit, Non-Reversible, Contour-Hugging MCMC

Wednesday 24
10:30 - 12:30
Room 7

Chris Sherlock[1] and Matthew Ludkin

[1]Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom
c.sherlock@lancaster.ac.uk

Both the Bouncy Particle Sampler (BPS) and the Discrete Bouncy Particle Sampler (DBPS) are non-reversible Markov chain Monte Carlo algorithms whose action can be visualised in terms of a particle moving across the statespace with some velocity. Both algorithms include an occasional step where the particle 'bounces' off a hyperplane which is tangent to the gradient of the target density, making the BPS rejection-free and allowing the DBPS to propose relatively large jumps whilst maintaining a high acceptance rate. Analogously to the concatenation of leapfrog steps in HMC, we describe an algorithm which omits the straight-line movement of the BPS and DBPS and, instead, at each iteration concatenates several discrete 'bounces' to provide a proposal which is on almost the same target contour as the starting point, producing a large proposed move which is very likely to be accepted. Combined with a separate kernel designed for moving between contours, an explicit bouncing scheme which takes account of the local Hessian at each bounce point ensures that the proposal respects the local geometry of the target, and leads to an efficient, skew-reversible MCMC algorithm.

Model Selection for Determinantal Point Processes

Wednesday 24
10:30 - 12:30
Room 5

Yasutaka Shimizu[1] and Kou Fujimori and Souta Sakamoto

[1]Applied Mathematics, Waseda University, Tokyo, Japan
shimizu@waseda.jp

The maximum composite likelihood estimators for stationary and isotropic parametric models of determinantal point processes (DPP) are discussed. Since the conditional marginal distribution of the point processes are given by determinant of positive definite kernels, we have the explicit form of the composite likelihoods for every order. This fact enables us to consider the generalized maximum composite likelihood estimator for every order. We will discuss the asymptotic properties using the limit theorems of DPPs. Based on those results, we will propose AIC-type information criteria for model selection.

Tuesday 23
16:30 - 17:30
Second Floor

New Advances in Multiple-Output Quantile Regression

Miroslav Šiman[1]

[1]Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 8, Czech Republic
siman@utia.cas.cz

Koenker's ordinary quantile regression has already been generalized to vector responses in a few ways. The poster introduces those extensions co-authored by the presenter (directional quantile regression, elliptical quantile regression, ameboid quantile regression) and highlights some recent results achieved in the development and application of these methods.

Monday 22
10:30 - 12:30
Room 10

Big Data and Functional Approach to Signal Analysis

Maria Skupień[1]

[1]Institute of Mathematics, Pedagogical University of Cracow, Kraków, Poland
marysia.skupien@gmail.com

Vibration signals sampled with a high frequency constitute a basic source of information about machine behaviour. Few minutes of signal observations easily translate into several millions of data points to be processed with the purpose of the damage detection. Big dimensionality of data sets creates serious difficulties with detection of frequencies specific for a particular local damage. In view of that, traditional spectral analysis tools like spectrograms should be improved to efficiently identify the frequency bands where the impulsivity is most marked (the so-called informative frequency bands or IFB). We propose the functional approach known in modern time series analysis to overcome these difficulties. We will process data sets as collections of random functions to apply techniques of the functional data analysis. As a result, we will be able to represent massive data sets through few real-valued functions and corresponding parameters, which are the eigenfunctions and eigenvalues of the covariance operator describing the signal. We will also propose a new technique based on the bootstrap resampling to choose the optimal dimension in representing big data sets that we process. Using real data generated by a gearbox and a wheel bearings we will show how these techniques work in practice.

Tuesday 23
16:30 - 17:30
Second Floor

A Comparison of Two Bayesian Accelerated Life Testing Models

Neill Smit[1] and Lizanne Raubenheimer and Thomas Mazzuchi

[1]School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa
neillsmit1@gmail.com

In this paper, two Bayesian accelerated life testing (ALT) models will be compared. High-reliability items are exposed to more severe than normal use conditions in ALT, in order to obtain failure data within a reasonable time frame. This failure data is then extrapolated to predict the life characteristics of the items in their normal functioning environment. The Birnbaum-Saunders and Weibull distributions are used as the life distributions, and the generalised Eyring model is incorporated as the time transformation function. This is a model with two stress variables, one thermal and one non-thermal. The models are used in an application, where Markov chain Monte Carlo (MCMC) methods are employed for posterior inferences.

Change-Point Modelling with Applications in Early Detection of Students at Risk

Thursday 25
10:30 - 12:30
Room 6

Georgy Sofronov[1] and Maurizio Manuguerra

[1]Department of Mathematics and Statistics, Macquarie University, Sydney, Australia
georgy.sofronov@mq.edu.au

In a large variety of applications, observations are taken sequentially over time and it is important to detect possible unpredictable changes in mechanisms of data generation. This problem is known as a change-point (or break-point) problem (see, for example, Priyadarshana, Polushina and Sofronov, 2015). The change-point problem arises in many different fields, including biomedical signal processing, speech and image processing, seismology, industry (e.g., fault detection) and financial mathematics. In this talk, we consider a sequential (or quickest) change-point problem where the data are observed sequentially (on-line) and our decision depends on the observations already made but does not depend on the future which is not yet known. We would like to detect the change-point as quickly as possible after it has occurred while the rate of false alarms is kept at a low predefined level. We propose a change-point model to analyse different students' pathways to graduation and to early detect students at risk (given various covariates). We introduce a notion of the standard student's pathway and calculate how each student's path differs from this standard. In this study, we develop a change-point detection model within the framework of transformation models (Manuguerra and Heller, 2018) and examine the student database at Macquarie University who have graduated with Science Degrees for the last 10 years.

References

Manuguerra, M., Heller, G.Z. (2018). ordinalCont: Ordinal Regression Analysis for Continuous Scales. R package version 2.0.0, URL https://cran.r-project.org/web/packages/ordinalCont/

Priyadarshana, M., Polushina, T. and Sofronov, G. (2015). Hybrid algorithms for multiple change-point detection in biological sequences. Advances in Experimental Medicine and Biology 823, 41–61.

Tuesday 23
10:30 - 12:30
Room 10

Noise Fit, Estimation Error and a Sharpe Information Criterion

Jakob Söhl[1] and Dirk Paulsen

[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands
j.soehl@tudelft.nl

When the in-sample Sharpe ratio is obtained by optimizing over a k-dimensional parameter space, it is a biased estimator for what can be expected on unseen data (out-of-sample). We derive (1) an unbiased estimator adjusting for both sources of bias: noise fit and estimation error. We then show (2) how to use the adjusted Sharpe ratio as model selection criterion analogous to the Akaike Information Criterion (AIC). Selecting a model with the highest adjusted Sharpe ratio selects the model with the highest estimated out-of-sample Sharpe ratio in the same way as selection by AIC does for the log-likelihood as measure of fit.

Wednesday 24
10:30 - 12:30
Room 9

Regression Analysis of Networked Data

Peter X. K. Song[1] and Yan Zhou

[1]Department of Biostatistics, University of Michigan, Michigan, United States
pxsong@umich.edu

We develop a new regression analysis approach to evaluating associations of covariates with outcomes measured from networks. This development is motivated from a study of infant growth that collects outcomes of event related potentials (ERP, a type of neuroimaging) measured over electroencephalogram (EEG) electrodes on the scalp. We propose a new generalized method of moments (GMM) that incorporates both established and data-driven knowledge of network topology among nodes in the estimation and inference to achieve robustness and efficiency. The GMM approach is computationally fast and stable to handle the regression analysis of network data, and conceptually it is simple with desirable properties in both estimation and inference. Both simulation studies and real EEG data analysis will be presented for illustration.

An Extension of the Censored Gaussian Lasso Estimator

Monday 22
10:30 - 12:30
Room 6

Gianluca Sottile[1] and Luigi Augugliaro and Veronica Vinciotti

[1]Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
gianluca.sottile@unipa.it

The conditional glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.

Astronomical Source Detection and Background Separation via Hierarchical Bayesian Nonparametric Mixtures

Thursday 25
10:30 - 12:30
Room 7

Andrea Sottosanti[1] and Mauro Bernardi and Roberto Trotta and David A. Van Dyk

[1]Statistical Sciences, University of Padova, Padova, Italy
sottosanti@stat.unipd.it

We propose an innovative approach based on Bayesian nonparametric methods to the signal extraction of astronomical sources in gamma-ray count maps under the presence of a strong background contamination. Our model simultaneously induces clustering on the photons using their spatial information and gives an estimate of the number of sources, while separating them from the irregular signal of the background component that extends over the entire map. From a statistical perspective, the signal of the sources is modeled using a Dirichlet Process mixture, that allows to discover and locate a possible infinite number of clusters, while the background component is completely reconstructed using a new flexible Bayesian nonparametric model based on b-spline basis functions. The resultant can be then thought of as a hierarchical mixture of nonparametric mixtures for flexible clustering of highly contaminated signals. We provide also a Markov chain Monte Carlo algorithm to infer on the posterior distribution of the model parameters which does not require any tuning parameter, and a suitable post-processing algorithm to quantify the information coming from the detected clusters. Results on different datasets confirm the capacity of the model to discover and locate the sources in the analysed map, to quantify their intensities and to estimate and account for the presence of the background contamination.

Friday 26
09:00 - 11:00
Room 7

Longitudinal Models with Informative Time Measurements

Inês Sousa[1] and Adriana Vieira

[1]Department of Mathematics, University of Minho, Guimarães, Portugal
isousa@math.uminho.pt

Longitudinal data analysis plays a key role in a multiple of distinct areas, including medicine. In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristics should allow for an extra information from the time measurements to the longitudinal process. In studies where observation time measurements and response variable are associated, a traditional longitudinal analysis will produce biased estimators and, consequently, uncertain conclusions. In this work we develop a simulation study to show how traditional longitudinal models perform under data generated with an association between time measurements and longitudinal response. It is also proposed a joint model for the distribution of longitudinal response and time measurement using maximum likelihood methodology to make inference on the model parameters. The model proposed is applied to a data set on progression of oncological biomarkers in breast cancer patients.

Monday 22
14:00 - 16:00
Room 6

Wavelet Regression Estimator for Compositional Data

Andrej Srakar[1]

[1]Institute for Economic Research (IER) and School of Economics and Business,
University of Ljubljana, Ljubljana, Slovenia
andrej_srakar@t-2.net

We provide a new nonparametric estimator for compositional data, based on generalization of the previous approach to any simplex space, study its statistical properties, and present simulation results of comparisons with existing estimators and two real data applications. Regression for compositional data has so far largely been considered only from a parametric point of view. Recently, some work adapted non-parametric regression to non-Euclidean manifolds, in particular the circular and spherical case. In a recent article, this approach was extended to compositional data situations, introducing local constant and local linear smoothing, treating the cases when either the response, the predictor or both of them are compositions. In our analysis, we extend their analysis to wavelet regressions, modeling the priors by use of wavelets constructed specifically for triangles, extending the results to any simplex space, to derive father and motherwavelets using special Legendre polynomial based sequential approach to orthogonalization. Finally, to derive the wavelet regression estimator, we refer to previous work studying Bayesian approach to regression with compositional data. We use multivariate wavelet priors, namely multivariate Laplace and multivariate Gaussian and study the derived regression estimator. The new estimator is derived for all three cases: simplicial-real; simplicial-simplicial; and real-simplicial regression. The preliminary results of simulation show that the new estimator outperforms the

more commonly used ones in both power and efficiency. Finally, we apply the findings to two economic datasets: SORS data for inference on inequality and UNCTAD database for inference on international trade. As compositional data regressions suffer from problems of adequate distributional features related to symplectic geometry, the article is a significant advance for the field and has a potential to be used frequently in future analyses.

Program Evaluation and Causal Inference for Distributional and Functional Data: Estimation of the Effects of Retirement on Health Outcomes

Monday 22
10:30 - 12:30
Room 10

Andrej Srakar[1]

[1]Institute for Economic Research (IER) and School of Economics, University of Ljubljana, Ljubljana, Slovenia
andrej_srakar@t-2.net

Statistical analysis of complex, i.e. non-standard data is gaining ground. Analysis of compositions, intervals, histograms, distributions and functions has become more and more common in contemporary statistics and econometrics. Despite several types of regressions existing for symbolic data, causal inference has not been studied so far adequately here. Furthemore, only slowly is it gaining ground using functional data. We develop a statistical theory for using instrumental variables with symbolic distributional data, related to the prevailing usage of regressions in such situations, the so-called "two components" Irpino and Verde model. We apply the findings to a pressing problem in the analysis of the aging process: the effects of retirement on health outcomes. Some authors conclude that retirement may lead to significant health improvements, but other studies find negative retirement effects. We use a panel dataset of Survey of Health, Ageing and Retirement in Europe (SHARE) in Waves 1-6. To address reverse causality in the relationship of retirement and health behaviours we use as instrument changes in retirement age. A novelty in the approach is that we treat countries as units and the variables are aggregated over countries. In this manner, we estimate the effect of the distribution of retirement across countries on distribution of health outcomes over countries (the instrumental variable is distributional as well). We are, therefore, able to estimate the causal effect for a group of countries (most commonly, exogenous change is used only to estimate the effect on one treated population). We extend the analysis for functional data and aggregate the variables to nonparametric functions, later used in functional linear models. As program evaluation and causal inference has so far not been studied with distributional data (and very seldom with functional data) the article is a significant step ahead in the analysis of complex data.

Monday 22
16:30 - 18:30
Room 5

Shapes Analysis of Functional Data

Anuj Srivastava[1]

[1]Department of Statistics, Florida State University, Tallahassee, FL, United States
anuj@stat.fsu.edu

Functional data has a growing presence in all branches of science and engineering, partly due to tremendous advances made in data collection and storage technologies. Such data is mostly analyzed using the classical Hilbert structure of square-integrable function spaces, but that setup ignores the shapes of these functions. Shape implies the ordering and the heights of peaks and valleys but is flexible on their exact locations. To focus on shapes of functions, we have introduced Elastic functional data analysis that allows time warpings of functions in order to register functional data, i.e. match their peaks and valleys. This, in turn, requires elastic Riemannian metrics that enable comparisons and testing of shape data modulo warping group action. I will present some statistical procedures resulting from their framework, including estimation of shape-constrained densities, ANOVA on shape space of curves, shape estimation and analysis of large biomolecules, and shape analysis of brain anatomical structures.

Friday 26
09:00 - 11:00
Room 8

Overlap Group Lasso in Functional Regression

Marco Stefanucci[1] and Antonio Canale and Mauro Bernardi

[1]Statistics, University of Padua, Padua, Italy
marco.stefanucci@unipd.it

Group Lasso methods are useful to shrink groups of coefficients to zero in a regression context. Sometimes it is useful to relate some of the variables to more than just one group, resulting in a model with an overlapping group structure. In this work we develop an Overlap Group Lasso regression model specifically tailored to induce sparsity in functional data. When the functional data are represented using a d-order B-spline system, a simple Lasso on the basis coefficients is not able to produce sparse estimates. Indeed a single zero in the coefficients of the basis functions is not enough to ensure a zero in the related function and one necessarily needs $d-1$ contiguous zeros. Our model groups together coefficients related to contiguous basis functions and for this reason is able to produce sparse estimates. To implement this approach, a novel Majorization-Minimization algorithm is proposed. This solution overcomes practical difficulties in the optimization due to the non-separability nature of the penalty function. Simulations and a stimulating real data application illustrate the performance of this approach.

A Sparse Beta Model with Covariates

Tuesday 23
16:30 - 17:30
Second Floor

Stefan Stein[1] and Chenlei Leng

[1]Statistics, University of Warwick, Coventry, United Kingdom
s.stein@warwick.ac.uk

The sparse β-model ($S\beta M$) is a new generative model for networks that interpolates the Erdős-Rényi model and the β-model by differentially modelling the parameters in the latter. To identify the support of and estimate the values of the parameters in the $S\beta M$, we propose using the penalized likelihood method with an ℓ_1 penalty. This immediately connects our estimation procedure to the LASSO approach which is popularly used for variable selection in regression. We exploit this connection to study the properties of the penalized estimator and provide numerical experiments to support the developed theory. In contrast to a penalized likelihood estimation method via the ℓ_0 penalty, we show that the use of the ℓ_1 penalty enables the incorporation of covariates while retaining the computational tractability as a convex optimization formulation, and discuss the related properties. This is joint work with Chenlei Leng.

Continuous-Time Locally Stationary Time Series Models

Tuesday 23
14:00 - 16:00
Room 6

Robert Stelzer[1] and Annemarie Bitter and Bennet Ströh

[1]Institute of Mathematical Finance, Ulm University, Ulm, Germany
robert.stelzer@uni-ulm.de

We adapt the classical definition of a sequence of locally stationary time series in discrete time to the continuous time setting using equivalent representations in the time and frequency domain. Special attention is given to time-varying state space processes, including the class of time-varying CARMA processes. Firstly the connection between these two classes of processes is examined and secondly suitable conditions on the coefficient functions are given, such that these processes are locally stationary. Finally, the Wigner-Ville is studied.

Varying-Sparsity Regression Models with Application to Cancer Proteogenomics

Tuesday 23
10:30 - 12:30
Room 9

Francesco Claudio Stingo[1] and Yang Ni and Veera Baladandayuthapani

[1]Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy
francescoclaudio.stingo@unifi.it

In this talk, I will describe recent approaches for the analysis of multi-omics data. Bayesian integration of heterogeneous data can facilitate the identification of patient-specific prognostic biomarkers, a critical step in the development of personalized treatment for clinically

and molecularly heterogeneous diseases such as cancer. The proposed methods allow flexible modeling of the biological interactions, as well as induces sparsity resulting in more parsimonious and interpretable models. Simulation studies demonstrate the superior performance of the proposed methods against competing method in terms of both marker selection and prediction. The application of the proposed methodology results in a better understanding of the underlying biological mechanisms.

Thursday 25
10:30 - 12:30
Room 10

Parameter Estimation of Finite Mixtures Based on the Empirical Identity Process

Ivo Stoepker[1] and Enrico Bibbona and Mauro Gasparini

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands
i.v.stoepker@tue.nl

When the empirical quantile function is composed with the empirical distribution function, we obtain a function that approximates the identity function. Appropriately referred to as the empirical identity function, we show that this function can be used to construct minimum-distance estimators for parametric settings. In this talk, we construct and study such estimators. We show that these estimators are critically dependent on the spacings in the data. This makes them of particular interest in parametric settings where we expect the structure of the spacings to vary considerably along the parameter space, such as finite normal mixtures. In particular, we show that they can outperform the EM-algorithm in the setting of finite normal mixtures, in the case of high component overlap.

Tuesday 23
10:30 - 12:30
Room 11

Models and Inference for On-Off Data via Clipped Ornstein–Uhlenbeck Processes

Emil A. Stoltenberg[1] and Nils Lid Hjort

[1]Department of Mathematics, University of Oslo, Oslo, Norway
emilas@math.uio.no

We study a model for recurrent event data subject to left-, intermittent- and right-censoring. The observations consist of binary sequences (along with covariates) for each individual under study. These sequences are modelled as generated by latent Ornstein–Uhlenbeck processes being above or below certain thresholds. Features of the latent process and the thresholds are taken as functions of covariates, allowing the researcher to distinguish factors that have an effect on the frailty, from those that have an effect on the variability of the observational unit. Inference is achieved by a quasi-likelihood approach, for which consistency and asymptotic normality is established. An advantage of our model is that particularities regarding the censoring need not be taken actively into account, and that it is well suited for situations where the individuals under study are irregularly and asynchronously observed. The motivation for our model came from a data set pertaining to

the incidence of diarrhoea among Brazilian children growing up under rather harsh conditions. We analyse these data with our model and contrast the results with intensity based counting process analysis of the same data.

Weak Dependence of Causal Random Fields and Statistical Applications

Monday 22
16:30 - 18:30
Room 7

Bennet Ströh[1] and Imma Valentina Curato and Robert Stelzer

[1]Institute of Mathematical Finance, Ulm University, Ulm, Germany
bennet.stroeh@uni-ulm.de

We introduce an easily accessible measure of dependence for random fields suitable for general concepts of causality (so called θ-lex-weak dependence), based on which we derive central limit theorems for stationary random fields. We show that for causal mixed moving average (MMA) fields driven by a Lévy basis satisfying this definition of weak dependence. As a consequence we give conditions ensuring that the sample mean and sample autocovariance are asymptotically normal. Finally, we apply the results to ambit fields and MSTOU processes.

Extremiles: A New Perspective on Asymmetric Least Squares

Thursday 25
14:00 - 16:00
Room 10

Gilles Stupfler[1] and Abdelaati Daouia and Irène Gijbels

[1]School of Mathematical Sciences, University of Nottingham, Nottingham, United Kingdom
gilles.stupfler@nottingham.ac.uk

Quantiles and expectiles of a distribution are found to be useful descriptors of its tail in the same way as the median and mean are related to its central behaviour. A valuable alternative class to expectiles, called extremiles, is considered; this class parallels the class of quantiles and includes the family of expected minima and expected maxima. The new class is motivated via several angles, which reveals its specific merits and strengths. Extremiles suggest better capability of fitting both location and spread in data points and provide an appropriate theory that better displays the interesting features of long-tailed distributions. We discuss their estimation in the range of the data and beyond the sample maximum. Some motivating examples are given to illustrate the utility of estimated extremiles in modelling noncentral behaviour. There is in particular an interesting connection with coherent measures of risk protection.

Monday 22
10:30 - 12:30
Room 11

Feature Estimation and Testing for Linear Regression with Time Series Regressors

Suhasini Subba Rao[1] and Raanju Sundararajan and Junho Yang

[1]Department of Statistics, Texas A&M University, Texas, USA
suhasini@stat.tamu.edu

There exists many real data examples, where the regressors of interest are high dimensional and come from a time series. For example, in econometric applications when dealing with multivariate models, the variables of interest are often observed at different time frequencies, such as financial and macroeconomic variables. The former being observable at high frequencies (e.g. daily, hourly or minute-by-minute) and the latter at far lower frequencies (often monthly or quarterly). Similarly in geostatistical applications it is important to understand how climatological data impacts climate change indicators, such as ice-shelf extent. Again, the climatological data (such as temperature) can be observed at a very high frequency, whereas the climate change indicators are often observed yearly. Consistent parameter estimation is usually only achieved with regularisation, which is usually done through dimension reduction or an additive penalty. But a disadvantage of most regularisation methods, is that the type of regularisation is tied to how the model is specified (smooth, sparse, periodic etc). Misspecification of the model can lead to spurious conclusions. However, if we treat the regressors as random variable, then the expectation of the normal equation will lead to a system of equations which are well posed (without the need to regularize). Therefore, in liue of any knowledge of the structure of the coefficients, the structure of the time series can be exploited to estimate the normal equations and thus consistently estimate the regression coefficients. In this talk, we propose a method for estimating the coefficients in a high dimension linear regression model, where the regressors come from a second order stationary time series. The proposed approach is based on deconvolution. This estimates, both the regression coefficients and its Fourier transform, which allows us to estimate different types of features in the regression coefficients. We show that the estimators are consistent and any finite subset are asymptotically normal. Further, we propose a method for estimating the asymptotic variance of coefficient estimators, which allows us to test the significance on the coefficients.

We apply our method to assessing the impact that daily temperatures have on the size of the Arctic ice shelf over the past 40 years.

Tuesday 23
10:30 - 12:30
Room 5

Estimation under Copula-Based Markov Mixture Normal Models for Time Series Data

Li-Hsien Sun[1] and Wei-Cheng Lin and Takeshi Emura

[1]National Central University, Taoyuan City, Taiwan
tpsun7246@gmail.com

We propose an estimation problem under a copula-based Markov model for serially correlated data. Motivated by the fat tail feature in stock market, we select a mixture normal distribution for the marginal distribution. Based on the mixture normal distribution for the marginal distribution and the Clayton copula for serial dependence, we obtain the corre-

sponding likelihood function. In order to obtain the maximum likelihood estimators, we apply the Newton Raphson method with appropriate transformations and initial values. In the em- pirical analysis, the stock price of Dow Jones Industrial Average is analyzed for illustration.

The X Factor: A Robust and Powerful Approach to X-Chromosome-Inclusive Whole-Genome Association Studies

Monday 22
14:00 - 16:00
Room 7

Lei Sun[1] and Bo Chen and Radu Craiu and Lisa Strug

[1]Department of Statisitical Sciences, University of Toronto, Toronto, Canada
sun@utstat.toronto.edu

The X-chromosome is often excluded from the so-called genome-wide association studies because of various analytical challenges. Some difficulties have been discussed previously, such as the random, skewed or no X-inactivation model uncertainty, while others have received little to no attention, such as the value in considering non-additive and gene-sex interaction effects, and the inferential consequence of choosing different baseline allele when analyzing the X-chromosome. In this talk we consider all these complexities jointly and propose a unifying inference framework based on regression. We provide theoretical justifications for its robustness in the presence various model uncertainties, as well as for its improved power under certain alternatives as compared with the existing approaches. For completeness, we also revisit the autosomes and show that the proposed framework leads to a robust and sometimes much more powerful test than the standard method. Finally, we provide supporting evidence from several application studies.

Valid Inference in Bandit Problems

Friday 26
09:00 - 11:00
Room 11

Yuekai Sun[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
yuekai@umich.edu

We consider the task of obtaining valid inferences from data collected by a regret-minimizing bandit algorithm. Recent work has shown that such data are biased, so naive inferences are invalid. We adapt the conditional approach to post-model selection inference to this task. As we shall see, this approach is well-suited for the task because it exploits randomization in the bandit algorithm to improve power.

Tuesday 23
10:30 - 12:30
Room 8

Asymptotic Distribution of the Bootstrap Parameter of an AR(p) Model

Bambang Suprihatin[1]

[1]Mathematics, Sriwijaya University, Indralaya, Indonesia
bambangs@unsri.ac.id

This paper is the extension of our research about asymptotic distribution of the bootstrap parameter estimator for the AR(1) model. We investigate the asymptotic distribution of the bootstrap parameter estimator of an AR(p) model by applying the delta method. The asymptotic distribution is the crucial property in inference of statistics. We conclude that the bootstrap parameter estimator of the AR(p) model asymptotically converges in distribution to the p-variate normal distribution.

Tuesday 23
16:30 - 17:30
Second Floor

Statistical Analysis of Stationary Fisher-Snedecor Diffusion

Nenad Suvak[1] and Nikolai N. Leonenko and Florin Avram and Ivan Papić

[1]Department of Mathematics, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia
nsuvak@mathos.hr

The problem of parameter estimation for a stationary diffusion with Fisher-Snedecor invariant distribution is observed. We propose moments based estimators of unknown parameters of invariant distribution and the autocorrelation parameter, based on both discrete and continuous observations, and prove their consistency and asymptotic normality. The explicit form of the asymptotic covariance matrix is determined by using the properties of eigenfunctions (Fisher-Snedecor polynomials) of the corresponding Sturm-Liouville operator. These results are derived in collaboration with Prof. N.N. Leonenko (School of Mathematics, Cardiff University, UK) and Prof. F. Avram (Department of Mathematics, University of Pau, France) and Ivan Papić (Department of Mathematics, J.J. Strossmayer University of Osijek, Croatia).

Thursday 25
14:00 - 16:00
Room 9

Stein Kernel Representations and their Applications

Marie Ernst, Gesine Reinert and Yvik Swan[1]

[1]Université de Liège, Liège, Belgium
yvik.swan@gmail.com

Stein kernels are central concepts introduced by Charles Stein in the 1970's. They provide new handles on probability distributions in terms of quantities whose first few moments suffice to characterize the underlying probability distributions. In this work we introduce

several generalizations, including extensions to arbitrary functionals of univariate distributions as well as kth order iterations. A new family of Fisher-type information quantities, directly inspired by this material, is also introduced. Several applications (ranging from the very theoretical to the very applied) are outlined.

PSICA: Decision Trees for Probabilistic Subgroup Identification with Categorical Treatments

Thursday 25
14:00 - 16:00
Seminari B

Oleg Sysoev[1] and Krzysztof Bartoszek and Eva-Charlotte Ekström and Katarina Ekholm Selling

[1]Computer Science, Linkoping University, Linkoping, Sweden
oleg.sysoev@liu.se

Personalized medicine aims at identifying best treatments for a patient with given characteristics. It has been shown in the literature that these methods can lead to great improvements in medicine compared to traditional methods prescribing the same treatment to all patients. Subgroup identification is a branch of personalized medicine which aims at finding subgroups of the patients with similar characteristics for which some of the investigated treatments have a better effect than the other treatments. A number of approaches based on decision trees has been proposed to identify such subgroups, but most of them focus on the two-arm trials (control/treatment) while a few methods consider quantitative treatments (defined by the dose). However, no subgroup identification method exists that can predict the best treatments in a scenario with a categorical set of treatments. We propose a novel method for subgroup identification in categorical treatment scenarios. This method outputs a decision tree showing the probabilities of a given treatment being the best for a given group of patients as well as labels showing the possible best treatments. The method is implemented in an R package **psica** available at CRAN. In addition to numerical simulations based on artificial data, we present an analysis of a community-based nutrition intervention trial that justifies the validity of our method.

Based on https://arxiv.org/abs/1811.09065

Spike and Slab Priors: Recovery, Uncertainty Quantification and Computational Issues

Wednesday 24
10:30 - 12:30
Room 12

Botond Szabo[1] and Ismael Castillo and Tim Van Erven and Kolyan Ray

[1]Mathematical Institute, Leiden University, Leiden, Netherlands
b.t.szabo@math.leidenuniv.nl

Spike and slab priors are frequently used Bayesian regularisation methods to induce sparsity in various high-dimensional models. In the first part of the talk I will investigate the theoretical properties (contraction rates and frequentist coverage of credible sets) of the corresponding posterior distribution. Then I will address the computational issues arising with the spike-and-slab prior. To overcome the high computational complexity we propose

a fast, deterministic algorithm to compute the posterior inclusion probabilities based on hidden Markov chains in context of the many normal means model. However, in context of the more general linear regression model one has to consider approximation methods. We show that (appropriately designed) mean-field variational Bayes methods can provide fast and optimal recovery of the underlying true signal.

Tuesday 23
14:00 - 16:00
Room 10

On Martingale CLT for Strictly Stationary Sequences

Zbigniew S. Szewczak[1]

[1]Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
zssz@mat.umk.pl

For strictly stationary martingale difference sequences we obtain complete characterization of the domain of attraction of the standard normal law (a'la Lévy, Khinchine and Feller). For this Rosenthal's type inequality is obtained and a new relative stability result.

Tuesday 23
14:00 - 16:00
Seminari B

Discrepancy Principle for Poisson Inverse Problems

Zbigniew Szkutnik[1]

[1]Faculty of Applied Mathematics, AGH University of Science and Technology, Krakow, Poland
szkutnik@agh.edu.pl

We consider the problem of inverse estimation of the intensity function of a Poisson process. Given observed a Poisson process (i.e. a random point measure) with the intensity function $g = Kf$, where K is a compact operator acting between some Hilbert spaces, we estimate the function of interest f, which is a form of an ill-posed Poisson inverse problem. Any regularization method requires a data-driven choice of some regularization parameter. We develope a general form of the Morozov discrepancy principle, suitable for such problems. Unlike the Morozov-type methods described in literature for Poisson data, we work in the infinite-dimensional setting. We present some nice asymptotic properties of our approach, as well as applications to some specific stereological problems.

Variational Inference for Stochastic Block Models from Sampled Data

Thursday 25
10:30 - 12:30
Room 11

Timothée Tabouy[1], Pierre Barbillon and Julien Chiquet

[1]MIA Paris, AgroParisTech, Paris, France,
timothee.tabouy@agroparistech.fr

This work deals with non-observed dyads during the sampling of a network and consecutive issues in the inference of the Stochastic Block Model (SBM). At first we will review sampling designs and recover Missing At Random (MAR) and Not Missing At Random (NMAR) conditions for the SBM. After that, we will introduce variants of the variational EM algorithm for inferring the SBM under various sampling designs (MAR and NMAR). We will also present model selection criteria based on Integrated Classification Likelihood for selecting both the number of blocks and the sampling design. In order to investigate the accuracy and the range of applicability of these algorithms we will show simulations. Finally, we will introduce SBM with covariates and show that different models without and with covariates, coupled with a sampling law giving rise to missing data of different natures, result in the same model.

On the Sign Recovery by LASSO, Thresholded LASSO and Thresholded Basis Pursuit Denoising

Tuesday 23
16:30 - 17:30
Second Floor

Patrik Tardivel[1] and Malgorzata Bogdan

[1]Institute of Mathematics, University of Wroclaw, Wroclaw, Poland
patrck.tardivel@gmail.com

We consider the high-dimensional regression model. In this presentation, we provide an upper bound for the probability of LASSO sign recovery which is reached when the non-null components of the true parameter are infinity large. In addition, when the irrepresentable condition holds, this upper bound provides a guide to select the LASSO's tuning parameter in order to control the FWER. On the other hand, LASSO can consistently estimate the true parameter under weaker assumptions than the irrepresentable condition. This implies that appropriately thresholded LASSO can recover the sign of the true parameter under such weaker assumptions. In this presentation we also revisit properties of thresholded LASSO and provide new theoretical results in the asymptotic setup under which the design is fixed and non-null components of the true parameter become infinity large. Apart from LASSO, our results cover also Basis Pursuit DeNoising (BPDN). Our main theorem takes a simple form: When non-null components of the true parameter are sufficiently large, appropriately thresholded LASSO or thresholded BPDN can recover the sign of the true parameter if and only if this parameter satisfies the identifiability condition. Finally, we illustrate how the knockoff methodology can be used to select an appropriate threshold and that thresholded BPDN and LASSO can recover the sign of the true parameter with a larger probability than adaptive LASSO.

Monday 22
16:30 - 18:30
Room 6

Optimal Change Detection Rules Maximizing Probability of Detection and their Application for Efficient Detection of Near-Earth Space Object Tracks

Alexander Tartakovsky[1] and Alexei Kolessa and Nikita Berenkov

[1]AGT Stat Consult and Moscow Institute of Physics and Technology, Los Angeles and Moscow, Russia
agt@phystech.edu

Changepoint problems deal with detecting changes in a process that occur at unknown points in time or space. A conventional approach to the sequential changepoint detection problem is to design a detection rule that minimizes the expected detection delay of a real change subject to a bound on the false alarm rate. However, in many applications, including space informatics, a more adequate optimality criterion is to maximize a probability of detection in a certain time interval subject to the constraint imposed on the false alarm risk. In this talk, we design optimal and nearly optimal change detection rules for Bayesian and minimax criteria. These rules are then applied to the problem of detection of near-Earth objects' tracks in images obtained by telescopes in the presence of noise and severe clutter from stars. The results show that even very weak tracks with unknown beginning and end can be effectively detected and localized.

Acknowledgement: The work was supported in part by the grant 18-19-00452 from the Russian Science Foundation, by the grant from the Russian Foundation for Advanced Research, and by the Russian Federation Ministry of Education and Science Arctic program and at the Moscow Institute of Physics and Technology.

Wednesday 24
10:30 - 12:30
Room 9

A Spatial Modeling Approach for Linguistic Object Data: Analysing Dialect Sound Variations Across Great Britain

Shahin Tavakoli[1] and Davide Pigoli and John Aston and John Coleman

[1]Department of Statistics, University of Warwick, Coventry, United Kingdom
s.tavakoli@warwick.ac.uk

Dialect variation is of considerable interest in linguistics and other social sciences. However, traditionally it has been studied using proxies (transcriptions) rather than acoustic recordings directly. We introduce novel statistical techniques to analyse geolocalised speech recordings and to explore the spatial variation of pronunciations continuously over the region of interest, as opposed to traditional isoglosses, which provide a discrete partition of the region. Data of this type require an explicit modeling of the variation in the mean and the covariance. Usual Euclidean metrics are not appropriate, and we therefore introduce the concept of d-covariance, which allows consistent estimation both in space and at individual locations. We then propose spatial smoothing for these objects which accounts for the possibly non convex geometry of the domain of interest. We apply the proposed method to data from the spoken part of the British National Corpus, deposited at the British Library, London, and we produce maps of the dialect variation over Great Britain. In addition, the methods allow for acoustic reconstruction across the domain of interest, allowing researchers to listen to the statistical analysis.

This is joint work with Davide Pigoli (King's College London), John Aston (Cambridge), and John Coleman (Oxford).

A Mean Score Equation-Based Approach to Correct for Nonignorable Verification Bias in Estimation of the Volume under the ROC Surface

Thursday 25
10:30 - 12:30
Seminari B

Duc-Khanh To[1] and Gianfranco Adimari and Monica Chiogna

[1]Department of Statistical Sciences, University of Padova, Padova, Italy
toduc@stat.unipd.it

The volume under the receiver operating characteristic surface (VUS) is used for measuring the overall accuracy of a diagnostic test when the possible disease status belongs to one of three ordered categories. In medical studies, the VUS of a new test is typically estimated through a sample of measurements obtained by some suitable sample of patients, for which the true disease status is assessed by a gold standard (GS) test. However, in many cases, due to the expensiveness and/or invasiveness of the GS test, only a subset of patients undergoes disease verification. Statistical inference based only on verified subjects is typically biased; this bias is known as verification bias. In order to correct for verification bias, the researchers must formulate some assumptions about the selection mechanism for the disease verification. When the decision to send a subject to verification may be directly based on the presumed subject's disease status, or the selection mechanism may depend on some unobserved covariates related to disease, the missing data mechanism is called nonignorable (NI). In this study, we propose a mean score approach to estimate the VUS under NI verification bias. We specify a parametric regression model for the verification process, which accommodates for possible NI missingness in the disease status. In order to avoid identifiability problems, we suggest to use instrumental variables. Then we consider a multinomial logistic model for the disease process based only on verified subjects, and estimate the model's parameters by solving mean score equations. By using the estimated verification and disease probabilities, we construct four NI bias-corrected VUS estimators. Consistency and asymptotic normality of the estimators are established. Simulation experiments are conducted to evaluate finite sample performances. An application to a dataset from a research on epithelial ovarian cancer is presented.

Scalable Inference for Epidemic Models with Individual Level Data

Monday 22
14:00 - 16:00
Room 10

Panayiota Touloupou[1] and Simon E.F. Spencer and Barbel Finkenstadt Rand

[1]Statistics, University of Warwick, Coventry, United Kingdom
P.Touloupou.1@warwick.ac.uk

As individual level epidemiological and pathogen genetic data become available in ever increasing quantities, the task of analysing such data becomes more and more challenging. Inferences for this type of data are complicated by the fact that the data are usually

incomplete, in the sense that the times of acquiring and clearing infection are not directly observed, making the evaluation of the model likelihood intractable. Data augmentation techniques implemented within the framework of Markov chain Monte Carlo (MCMC) methods can tackle these problems by taking into account the unobserved dynamics of transmission and thus have been widely employed in practice.

Until recently, the study of infectious diseases in large scale populations has been challenging due to the computationally intensive methods needed to these models. In this work we describe a novel method for updating individual level infection states within MCMC algorithms that respects the dependence structure inherent within epidemic data. Our algorithm achieves a good balance between computational complexity and mixing properties and therefore can be viewed as an alternative to existing approaches, particularly for applications with large populations.

We apply our new methodology to an epidemic of Escherichia coli O157:H7 in feedlot cattle in which 8 competing strains were identified using genetic typing methods. We show that surprisingly little genetic data is needed to produce a probabilistic reconstruction of the epidemic trajectories, despite some possibility of misclassification in the genetic typing. We believe that this complex model, capturing the interactions between strains, would not have been able to be fitted using existing methodologies. Joint work with Simon Spencer and Bärbel Finkenstädt Rand.

Tuesday 23
16:30 - 17:30
Second Floor

Model-Based Approach in Dissolution Profile Comparison: Alternative to F2 Metric

Olympia Tumolva[1] and Jean-François Michiels

[1]Pharmalex Belgium, Mont-Saint-Guibert, Belgium
olympia.tumolva@pharmalex.com

Dissolution profile comparison provides a valuable information for pharmaceutical product quality when post-approval changes have been made to the product. These changes include modifications in formulation, changes to the manufacturing process or the site of manufacture, and in process scale-up. The f2 metric is widely used for comparing dissolution profiles between a test (pre-changed batches) and reference (post-changed batches) group. However, its use is restrictive and drug authorities proposed alternatives. Thus, this paper explores two methodologies on dissolution profile comparison for a pharmaceutical product. First, a similarity region is defined based on the variation of parameters of a multivariate model that was fitted on resampled sets of data from the reference group. The multivariate model is fitted on the data of reference and test group and the parameters of this model were checked if they are within the defined similarity region. Second, a Bayesian predictive approach is used wherein a Bayesian multivariate model was fitted separately to the data of the test and reference group. Posterior predictive distribution (PPD) is generated for the reference and test group. Tolerance interval (TI) is constructed from the PPD of the reference group and the probability that the generated observations of the test group are within the TI is determined. Illustration and comparison of the methods will be done thru simulated and real-life data.

Tailoring PCA for Detecting Sparse Changes in Multi-Stream Data

Thursday 25
10:30 - 12:30
Room 6

Martin Tveten[1] and Ingrid K. Glad

[1]Mathematical Department, University of Oslo, Oslo, Norway
martintv@math.uio.no

Motivated from high-dimensional data stream or sensor network applications, we study the problem of detecting changes in the mean and/or the covariance matrix of normally distributed data in an online/sequential manner. Special attention is given to the case of sparse changes, i.e., when only a few streams are affected by a change. The aim is to detect such changes as quickly as possible, while controlling false alarms.

To make the procedure scalable, our strategy of interest is based on projecting the incoming data onto a few of the principal axes of the pre-change data. We present an algorithm that tailors the choice of projections to monitor when change detection is the goal. The tailoring is based on a training set of normal-state observations as well as a possible specification of which change scenarios that are of main interest. Most often, the least varying projections are selected, and we provide some theoretical justification for this. A combination of tailored PCA and a generlized log-likelihood monitoring procedure displays high efficiency in detecting even very sparse changes in the mean, variance and correlation, in simulations. Simultaneously, the computational burden is reduced considerably.

Parametric Inference for a Discretely Observed SPDE

Wednesday 24
10:30 - 12:30
Room 5

Masayuki Uchida[1] and Yusuke Kaino

[1]Osaka University, Osaka, Japan
uchida@sigmath.es.osaka-u.ac.jp

We consider parametric estimation of a parabolic linear second order stochastic partial differential equation (SPDE) from ultra high frequency data. Markussen (2003, Bernoulli) investigated quasi maximum likelihood estimators of unknown parameters in a SPDE based on sampled data whose discretization step is fixed and whose sample size tends to infinity. Bibinger and Trabs (2017, arXiv:1710.03519) studied least squares type estimators of the normalized volatility parameter and the curvature parameter in a SPDE from high frequency data. In this talk, we derive adaptive estimators of the coefficient parameters in the SPDE by using the estimators of the normalized volatility parameter and the curvature parameter. In order to get the adaptive estimators, we used the thinned data obtained from the ultra high frequency data. For details of adaptive estimators and thinned data, see Uchida and Yoshida (2012, SPA) and Kaino and Uchida (2017, SISP), respectively. Furthermore, examples and simulation results are given.

Thursday 25
10:30 - 12:30
Room 5

Noise Estimation for Ergodic Levy Driven SDE in YUIMA Package

Yuma Uehara[1] and Hiroki Masuda and Lorenzo Mercuri

[1]The Institute of Statistical Mathematics, Tokyo, Japan
y-uehara@ism.ac.jp

Levy driven stochastic differential equation (SDE) is an important building block for modeling non-Gaussian high-frequency data observed in many fields such as biology, life sciences, engineering, pharmacodynamics, and so forth. It has been, however, a longstanding problem that a closed form of the likelihood function is rarely available except for quite limited special cases, so that, in particular, information of the driving Levy noise is difficult to be estimated from observed data. In this talk, we propose a multistep estimation procedure based on the Euler residuals constructed from the Gaussian quasi-maximum likelihood estimator: we approximate unit time increments of the driving noise by partially summing up the Euler residuals, and then apply parametric estimation methods. We will present theoretical property of the proposed estimator and numerical experiments through YUIMA package in R. This is a joint work with Hiroki Masuda (Kyushu University) and Lorenzo Mercuri (University of Milan).

Tuesday 23
09:00 - 10:00
Aula Magna
Building 13

Nonparametric Bayes: Review and Challenges

Aad van der Vaart[1]

[1] Mathematical Institute, Leiden University, Leiden, Netherlands,
avdvaart@math.leidenuniv.nl

Nonparametric Bayesian methods have seen a great development in the past decades. They are ordinary Bayesian methods using a prior distribution on an infinite-dimensional or high-dimensional parameter (function, distribution, high-dimensional regression vector), resulting in a posterior distribution, giving the plausibility of this parameter given the data. Nonparametric Bayesian methods are now routinely applied in many areas, based on the promise of reconstruction, through the mode or mean of the posterior distribution, and automatic uncertainty quantification, through the spread in the posterior distribution. Besides from statisticians they attract attention of computer scientists and mathematical analysts, in particular in connection with inverse problems and data assimilation. Through empirical Bayes ideas they are connected to the 'sharing of information' and large scale inference in settings of high-dimensional data. There is an increasing theoretical understanding of the performance of these methods from the non-Bayesian perspective, developed under the assumption that the prior is only a working hypothesis to model a true state of nature. Theory has been developed for classical nonparametric smoothing problems, sparse high-dimensional models and increasingly for inverse problems, and addresses a great variety of priors, based on the classical Dirichlet process, Gaussian processes, spike-and-slab distributions, and many others. One of the attractions is the automatic adaption to complexity by means of hyperparameters fitted through hierarchical and empirical Bayes approaches. Theory addresses rates of contraction of the posterior to a true parameter, distributional approximations of the posterior distribution of smooth functionals, and most

recently the coverage (or lack of it) of Bayesian credible sets. In this talk we present some examples of success stories, and point to open questions.

An Empirical Bayes Approach to Co-Data Learning in Ridge Models

Tuesday 23
10:30 - 12:30
Room 9

Mirrelijn Van Nee[1] and Mark Van De Wiel

[1]Amsterdam UMC, VU, Amsterdam, Netherlands
m.vannee@vumc.nl

Prediction in high-dimensional settings, in which the number of covariates exceeds the number of samples, is hard. However, often there's auxiliary data on the covariates (co-data) available which could be used to improve predictions. The Empirical Bayes (EB) approach takes advantage from the high-dimensionality by using the data to estimate shrinkage prior parameters, or equivalently, regularisation penalties. We present an EB method with which we are not only able to include co-data to learn multiple prior parameters for different covariates, but in which we can also add an additional layer of shrinkage to account for structure of the co-data.

More specifically, we consider generalised ridge regression, in particular binary and survival. We use co-data to define possibly overlapping or hierarchical covariate groups that may differ considerably in terms of predictive strength, such that the prediction would potentially benefit from penalising these groups by different ridge penalties. Our proposed method is an efficient moment-based EB approach to find estimates for the group penalties. The extra layer of shrinkage enables inclusion of the known structure of overlapping and hierarchical groups, and leads to stable group penalty estimates for large numbers of groups. Any type of shrinkage, e.g. ridge or hierarchical lasso, can be used in this layer, rendering a flexible framework to improve predictions. Also, by using hierarchical lasso in the extra layer, the method adequately handles continuous co-data by estimating the relation between the penalties and continuous covariates non-parametrically. Moreover, the framework allows for integration and weighting of multiple co-data sets, as well as posterior variable selection. As an illustrating example, we demonstrate the method in an oncogenomics setting.

Detection of Communities in the Stochastic Block Model: Consistency and Confidence Sets

Tuesday 23
16:30 - 17:30
Second Floor

Jan Van Waaij[1] and Bas Kleijn

[1]Statistics institute, University of Padova, Padova, Italy
jvanwaaij@gmail.com

Posterior distributions for community assignment in the stochastic block model are shown to achieve frequentist exact recovery and detection under sharp lower bounds on sparsity. Assuming posterior recovery (or detection), one may interpret credible sets (or enlarged credible sets) as consistent confidence sets.

Thursday 25
10:30 - 12:30
Room 9

Solving Estimating Equations with Copulas

Thibault Vatter[1] and Thomas Nagler

[1]Department of Statistics, Columbia University, New York, United States
thibault.vatter@columbia.edu

Thanks to their ability to capture complex dependence structures, copulas are frequently used to glue random variables into a joint model with arbitrary one-dimensional margins. More recently, they have been applied to solve statistical learning problems such as regression or classification. Framing such approaches as solutions of estimating equations, we generalize them in a unified framework. We derive consistency, asymptotic normality, and validity of the bootstrap for copula-based Z-estimators. We further illustrate the versatility of such estimators through theoretical and simulated examples.

Monday 22
16:30 - 18:30
Room 12

Detecting the Direction of High-Dimensional Spherical Signals

Thomas Verdebout[1] and Davy Paindaveine

[1]ECARES and Mathematics department, Université libre de Bruxelles, Bruxelles, Belgium
tverdebo@ulb.ac.be

We consider one of the most important problems in directional statistics, namely the spherical location testing problem, whose null is that the modal location $\boldsymbol{\theta}$ of a Fisher–von Mises–Langevin (FvML) distribution on the p-dimensional unit sphere $\mathcal{S}^{p-1}$ coincides with a given location $\boldsymbol{\theta}_0$. The underlying concentration parameter κ plays the role of a nuisance. We derive local asymptotic normality (LAN) results in a general high-dimensional framework where the dimension $p = p_n$ goes to infinity, at an arbitrary rate, with the sample size n, and where the concentration κ_n behaves in a completely free way with n, which offers a spectrum of problems ranging from arbitrarily easy to arbitrarily challenging ones. We identify seven asymptotic regimes, depending on the convergence/divergence properties of (κ_n), that yield different limiting experiments and different contiguity rates. In each regime, we derive Le Cam optimal tests and we compute, from the Le Cam third lemma, asymptotic powers of the classical Watson test under contiguous alternatives. To obtain a full understanding of the non-null behavior of this test, we derive its local asymptotic powers in the broader, semiparametric, model of rotationally symmetric distributions. Monte Carlo studies show that finite-sample behaviours remarkably agree with our asymptotic results.

Bias Reduced Simulation-Based Estimators in High Dimensions

Tuesday 23
14:00 - 16:00
Room 12

Maria-Pia Victoria-Feser[1] and Stephane Guerrier and Samuel Orso and Mucyo Karemera

[1]Research Center for Statistics, University of Geneva, Geneva, Switzerland
maria-pia.victoriafeser@unige.ch

Considering the increasing size of available data, the need for statistical methods that control bias, while remaining computationally tractable, is growing. This is mainly due to frequent high dimensional settings and the increasing complexity of statistical models and estimators that bring many inferential procedures to incur significant loss in terms of statistical and computational performance. In this paper, we propose a general framework from which estimators can be derived in a computationally efficient manner for a wide class of models. It requires an initial estimator, possibly baised and non consistent, and the final estimator is obtained by using the Iterative Bootstrap (Guerrier, et al, 2018, 2019) which converges exponentially fast. This simulation-based framework is particularly well suited for estimators that are robust to spurious data contamination, for large discrete and/or non Gaussian models with a large number of parameters. Indeed, a simple weighted MLE (non consistent), for example, can be used as initial estimator, avoiding therefore the approximation of multiple intergrals. It is also convenient for models for which the maximum likelihood estimator is either to complex to implement or simply does not exist. Within this framework we prove the properties of simulation-based estimators, more specifically the conditions for asymptotic unbiasedness, consistency and asymptotic normality when the number of parameters is allowed to increase with the sample size. In a simulation study covering several data and model settings, we find evidence of the advantages of our simulation based estimator in terms of finite sample mean squared error of estimation over other available estimators.

Inference for Multiplicative Model Combination using Score Matching

Wednesday 24
10:30 - 12:30
Seminari B

Paolo Vidoni[1]

[1]Department of Economics and Statistics, University of Udine, Udine, Italy
paolo.vidoni@uniud.it

This paper concerns multiplicative model combination, which is a way of combining probability models using a weighted multiplication and a subsequent normalization. In particular, we focus on density estimation problems and we define a density estimator, based on a suitable model combination, using a new boosting-type algorithm with the Hyvärinen score as loss function. Finally, a simple application to the estimation of the precision matrix of a multivariate Gaussian model is presented.

Thursday 25
14:00 - 16:00
Room 10

Extreme Value Theory for Open Set Classification - GPD and GEV Classifiers

Edoardo Vignotto[1] and Sebastian Engelke

[1]University of Geneva, Geneva, Switzerland
edoardo.vignotto@unige.ch

Classification tasks usually assume that all possible classes are present during the training phase. This is restrictive if the algorithm is used over a long time and possibly encounters samples from unknown classes. It is therefore fundamental to develop algorithms able to distinguish between known and unknown new data. In the last few years, extreme value theory has become an important tool in multivariate statistics and machine learning. This is due to the fact that the extreme features, rather than the average ones, are often most important for discriminating between different objects. The recently introduced extreme value machine, a classifier motivated by extreme value theory, addresses this problem and achieves competitive performance in specific cases. This algorithm strongly relies on the knowledge about the known classes to infer about the unknowns and gives a non-justified premium to known classes far from all the others. We show that the extreme value machine can thus fail when the geometries of known and unknown classes differ, even if the recognition problem is fairly simple. To overcome these limitations, we propose two new algorithms for open set classification that share similar foundations but are more robust in such cases. We exploit the intuition that test points that are extremely far from the training classes are more likely to be unknown objects. We derive asymptotic results motivated by univariate extreme value theory that make this intuition precise. The effectiveness of our classifiers is illustrated in simulations and on real data sets. Finally, we state their main strengths and some possible future improvements.

Thursday 25
10:30 - 12:30
Room 10

Graph Inference with Clustering and False Discovery Rate Control

Fanny Villers[1] and Tabea Rebafka and Etienne Roquain

[1]Sorbonne Université, Paris, France
fanny.villers@upmc.fr

Networks analysis is widely used to model and describe interactions between entities. In many situations, the practioner does not directly observe the interaction graph, but only a noisy version of it. This is for example the case of similarity graphs or when edges represent correlations between the nodes. To recover the underlying graph, ad hoc methods such as thresholding rules are often used, which can cause a degradation of the topology of the network and may lead to wrong interpretations, especially when combined with cluster analysis.

In this work, we introduce a noisy version of the stochastic block model and propose to recover the underlying graph with a false discovery rate (FDR) guarantee, that is, by controlling the average proportion of errors among the edges declared as significant. A novelty of our method is that the graph and the clustering are inferred simultaneously, so that the two types of inference can help each other. Our methodology relies upon

Bayesian multiple testing techniques with parameters estimated by a suitable variational EM algorithm. The performances of our method are supported both by theoretical findings and by numerical experiments.

Weighted Approximate Bayesian Computation via Sanov's Theorem

Tuesday 23
10:30 - 12:30
Room 6

Cecilia Viscardi[1] and Michele Boreale and Fabio Corradi

[1]Dipartimento di Statitstica, Informatica, Applicazioni "G.Parenti", University of Florence, Firenze, Italy
cecilia.viscardi@unifi.it

Approximate Bayesian Computation (ABC) is a widespread likelihood-free method which allows simulated inference on model parameters. The method only requires the availability of simulations from the model and it is not influenced by the sometime intractable form of the likelihood. This method provides a conversion of samples from the prior parameters distribution into samples from the posterior by rejecting those parameters producing simulated observations very different from the observed data – i.e. at a distance greater than a specified threshold. Thus, parameter values which rarely lead to simulations resembling the observed data are often discarded altogether. In certain situations, the vast majority of parameters proposals is "poor" – i.e. has an (exponentially) small probability of producing simulation outcomes close to the observed data – and thus discarded. As such, they do not contribute at all to the representation of the parameters posterior distribution. This typically leads to a huge number of simulations and/or a waste of computational resources. Our idea is to resort to the Large Deviations Theory to evaluate the probability of such rare events – a poor parameter proposal producing an outcome close to the observed data – with the aim to provide a better approximation of the tails of the parameters posterior distribution. Accordingly, we propose a Weighted Approximate Bayesian Computation algorithm (W-ABC) where parameters weights are computed via the Sanov's Theorem thus avoiding the rejection step. Here, in order to derive a computable asymptotic approximation for the Sanov's result, we adopt the Method of Types, thus restricting our methodology to models for discrete random variables. In addition, we deal with the scalability of the W-ABC w.r.t. the size of the sample space by investigating different choices for the required distance function. Finally, we evaluate the performances of our methodology through a proof-of-concept implementation.

Thursday 25
10:30 - 12:30
Room 7

The DAB: Detecting Anomalies via Bootstrapping

Pietro Vischia[1] and Tommaso Dorigo and Giovanna Menardi and Grzegorz Kotkowski

[1]Institut de recherche en mathématique et physique, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
pietro.vischia@cern.ch

We develop an algorithm relying on the idea of testing how likely experimental data have been generated by the background distribution, i.e. under the null hypothesis of the absence of any anomaly; bootstrap procedures are applied to build a set of scores aiming to classify events as background-like or signal-like. We extend the study of the algorithm performance in simple uni- and multivariate scenarios to confirm our original conjectures (arXiv:1611.08256), and explore alternative choices for the inner structure of the algorithm. Finally, we study the application to realistic scenarios from Particle Physics.

Monday 22
16:30 - 18:30
Room 8

2D Real Microstructures Images: a Great Source of Data

Martina Vittorietti[1] and Geurt Jongbloed and Jilt Sietsma and Wei Li

[1]Applied Mathematics, Delft University of Technology, Delft, Netherlands
m.vittorietti@tudelft.nl

Modeling microstructures is an interesting problem not just in Materials Science but also in Mathematics and Statistics. Moreover, understanding the relationship between microstructure characteristics and mechanical properties of materials is not a trivial problem. Images of 2D section of metals microstructure are the main source of data available for understanding this kind of problems. In this work, we show how to treat the data in two different cases: a testing framework for Poisson-Voronoi model assumption and a modeling framework for explaining how the presence of carbides in metals microstructure can influence the mechanical behavior of the material.

Thursday 25
10:30 - 12:30
Room 9

Estimating the Extremal Dependence Structure of Time Series Extremes using Block Maxima

Stanislav Volgushev[1] and Nan Zou and Axel Bücher

[1]Department of Statistics, University of Toronto, Toronto, Canada
stanislav.volgushev@utoront.ca

In this talk we introduce sliding block methods for estimating the extremal dependence structure (tail copula) of a vector-valued time series. We derive functional central limit theorems for empirical copulas of block maxima that are uniform in the block size parameter, which seems to be the first result of this kind for estimators based on block maxima in general. The theory allows for various aggregation schemes over multiple block sizes,

leading to substantial improvements over the single block length case and opens the door to further methodology developments. In particular, we consider bias correction procedures that can improve the convergence rates of extreme-value estimators and shed some new light on estimation of the second-order parameter when the main purpose is bias correction.

Intrinsic Wavelet Regression for Curves and Surfaces of Hermitian Positive Definite Matrices

Friday 26
09:00 - 11:00
Room 12

Rainer Von Sachs[1]

[1]ISBA/LIDAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
rvs@uclouvain.be

In multivariate time series analysis, non-degenerate autocovariance and spectral density matrices are necessarily Hermitian and positive definite. An estimation methodology which preserves these properties is developed based on intrinsic wavelet transforms being applied to nonparametric wavelet regression for curves in the non-Euclidean space of Hermitian positive definite matrices. Via intrinsic average-interpolation in a Riemannian manifold equipped with a natural invariant Riemannian metric, we derive the wavelet coefficient decay and linear wavelet thresholding convergence rates of intrinsically smooth curves. Applying this more specifically to nonparametric spectral density estimation, an important property of the intrinsic linear or nonlinear wavelet spectral estimator under the invariant Riemannian metric is that it is independent of the choice of coordinate system of the time series, in contrast to most existing approaches. In this talk, we give insights, discuss some theoretical results and present some numerical examples of our approach, and we also mention the generalisation of this one-dimensional denoising of matrix-valued curves in the Riemannian manifold to higher-dimensional intrinsic wavelet transforms. This talk is based on joint work with Joris Chau.

Change-Point Detection Based on Weighted Two-Sample U-Statistics

Tuesday 23
16:30 - 17:30
Second Floor

Kata Vuk[1] and Herold Dehling and Martin Wendler

[1]Fakultät für Mathemati, Ruhr-Universität Bochum, Bochum, Germany
kata.vuk@rub.de

We consider a robust change-point test which is based on weighted two-sample U-statistics. We focus on short range dependent data, more precisely on data that can be represented as functionals of a mixing process. In this way, most examples from time series analysis are covered. Under the hypothesis that no change occurs, the limit distribution of the the considered test statistic is derived. Under the alternative of a change-point with constant hight, we derive consistency. The considered test is sensitive on tails, which means that early and late changes can be better detected. To illustrate the results and to investigate the power of the test, we will give some simulation results.

Monday 22
16:30 - 18:30
Room 7

Identification of the Number of Factors for Factor Modeling in High Dimensional Time Series

Qinwen Wang[1]

[1]School of Data Science, Fudan University, Shanghai, China
wqw@fudan.edu.cn

Identifying the number of factors in a high-dimensional factor model has attracted much attention in recent years and a general solution to the problem is still lacking. A promising ratio estimator based on the singular values of the lagged autocovariance matrix has been recently proposed in the literature and is shown to have a good performance under some specific assumption on the strength of the factors. Inspired by this ratio estimator and as a first main contribution, we will propose a complete theory of such sample singular values for both the factor part and the noise part under the large-dimensional scheme where the dimension and the sample size proportionally grow to infinity. In particular, we provide the exact description of the phase transition phenomenon that determines whether a factor is strong enough to be detected with the observed sample singular values. Based on these findings, we propose a new estimator of the number of factors which is strongly consistent for the detection of all significant factors (which are the only theoretically detectable ones). In particular, factors are assumed to have the minimum strength above the phase transition boundary which is of the order of a constant; they are thus not required to grow to infinity together with the dimension (as assumed in most of the existing papers on high-dimensional factor models).

Tuesday 23
16:30 - 17:30
Second Floor

CLT for Linear Spectral Statistics of a Class of Wigner-Type Matrices with General Variance Profiles

Zhenggang Wang[1] and Zhigang Bao and Jeff Yao

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
u3005312@connect.hku.hk

Hypothesis testing on certain random graph models such as ER Model and Stochastic Block Model often proceeds via spectral properties of the adjacency matrix or graph Laplacian, namely the fluctuations of the eigenvalues and eigenvectors of the associated matrices. Though fluctuations of the local statistics are widely studied and known, classical results on the fluctuations of global eigenvalue statistics are often limited to the homogeneous case. This paper investigates the linear spectral statistics of a class of Wigner-type matrices with a general variance matrix, i.e. with possibly nonhomogeneous second and fourth moments. We establish a CLT for the linear spectral statisitcs with the resolvent method, which allows us to propose a class of test statistics for the nonhomogeneous stochastic block model with comparable but not necessarily equal-sized communities.

Evaluate Association and Diagnosis of Disease using Longitudinal Markers with Application to Alzheimer's Disease Neuropathology

Tuesday 23
16:30 - 17:30
Second Floor

Cuiling Wang[1] and Ellen Grober and Richard Lipton

[1]Albert Einstein College of Medicine, Bronx, United States
cuiling.wang@einstein.yu.edu

In epidemiological and clinical studies, it is often important to evaluate the association of a marker with a disease and how it can be used for the diagnosis of disease. In longitudinal studies, repeated measurements of markers are often available. Evaluating the association of ante-mortem cognitive performance with post-mortem AD neuropathology is critically important for understanding the disease and the early detection of AD so that effective intervention can be applied. Missing data due to study design and/or drop out often arise as a major challenge in this analysis. The commonly used approach in practice of arbitrarily selecting one time-point of measure (e.g., at last visit) is convenient but makes strong assumptions and loses information. The appropriate statistical approaches include multiple imputation and joint modeling. We propose to model the trajectory of the longitudinal marker given the disease to answer the questions on the association as well as ROC related questions. The method is applied to an example of using the longitudinal free recall score of the Free and Cued Selective Reminding Test (FCSRT) with Immediate Recall to diagnose Alzheimer's Disease (AD) neuropathology in the Einstein Aging Study (EAS). Simulation studies were performed to evaluate the performance of our methods.

Essential Regression

Monday 22
14:00 - 16:00
Room 12

Marten Wegkamp[1] and Mike Bing and Florentina Bunea and Seth Strimas-Mackey

[1]Mathematics, Cornell, Ithaca, United States
mhw73@cornell.edu

We introduce the Essential Regression model, which provides an alternative to the ubiquitous K-sparse high dimensional linear regression on p variables. While K-sparse regression assumes that only K components of the observable X directly influence Y, Essential Regression allows for all components of X to influence Y, but mediated through a K-dimensional random vector Z. The mediator Z is unobservable, and made interpretable via a modeling assumption through which each component of Z is given the physical meaning of a small group of the X-variables. Formally, E-Regression is a new type of latent factor regression model, in which the unobservable factors Z influence linearly both the response Y and the covariates X. Its novelty consists in the conditions that give Z interpretable meaning as well as render the regression coefficients β relating Y to Z – along with other important parameters of the model – identifiable. Inference performed in regression at the lower resolution level given by Z is uniform over the space of the regression coefficients β, although it is performed after estimating consistently K and the subset of the X-variables that explain Z. We construct computationally efficient estimators of β, derive their minimax rate, and show that they are minimax-rate optimal and asymptotically normal. Prediction of Y from X under E-Regression complements, in the low signal to noise ratio regime, the battery

of methods developed for prediction under other factor model specifications. Similarly to other methods, it is particularly powerful when p is large.

Friday 26
09:00 - 11:00
Room 9

Statistical Inference for Molecules: How many and Where?

Frank Werner[1] and Katharina Proksch and Jan Keller-Findeisen and Axel Munk

[1]Institute for Mathematical Stochastics, University of Goettingen, Göttingen, Germany
f.werner@math.uni-goettingen.de

Nowadays it is possible to image fluorescent probes on a sub-diffraction spatial resolution due to super-resolving scanning microscopes like STED. Nevertheless, none of the available techniques provides information about the distribution of molecules, but only about the total brightness distribution. As the total brightness is the product of local brightness and molecule number, it is a key problem to access the local brightness and infer on the molecule distribution. In this talk we present a sound statistical model which connects both the local brightness and molecule distribution with the coupled probabilities to simultaneously detect k photons at the same spatial coordinate and time (k = 1, 2, 3, 4). Using a modified fluorescence microscope approximations of these quantities can be measured in practice. The resulting equations give rise to a nonlinear statistical inverse problem, which allows to identify the local brightness and molecule distribution from observable quantities. However, solving it remains challenging, and furthermore it is difficult to quantify the uncertainty in corresponding reconstructions. We present a hybrid approach which tackles inferring on the number and location of molecules in the sample separately, but still allows to obtain uniform confidence statements about the resulting molecular map.

Tuesday 23
10:30 - 12:30
Room 6

Dynamic Visualization and Fast Computation for Convex Clustering via Algorithmic Regularization

Michael Weylandt[1] and John Nagorski and Genevera Allen

[1]Rice University, Houston, United States
michael.weylandt@gmail.com

We introduce algorithmic regularization, a novel approach to efficient computation of the entire solution path of penalized estimators using an iterative one-step approximation scheme. The effectiveness of algorithmic regularization is illustrated with an extended application to convex clustering, where it achieves over a 100-fold speed-up over existing methods while constructing a finer approximation grid than standard methods. These improvements allow creation of a convex-clustering dendrogram and dynamic path-wise visualizations based on modern web technologies. We justify our approach with a novel theoretical result, guaranteeing global convergence of the approximate path to the exact path under easily-checked non-data-dependent assumptions. Finally, we will discuss the application of algorithmic regularization to other problems in machine learning, and discuss some future directions. Our methods are implemented in the open-source R package clustRviz, available at https://dataslingers.github.io/clustRviz. This talk is based on joint work with Genevera Allen, John Nagorski, and Yue Hu.

Probabilities of Exceeding Mean or Mode: Ordering the Beta-Distributions by Skewness

Friday 26
09:00 - 11:00
Seminari B

Tilo Wiklund[1] and Idir Arab

[1]Department of Mathematics, Uppsala University, Uppsala, Sweden
tilo.wiklund@math.uu.se

There is of course a very well developed theory (concentration of measure, large deviations, ...) concerning probabilities of deviating from a measure of location (expectation, median, mode). Less appears to be known about probabilities of exceeding a measure of location, that is to say deviating in certain *direction* by an arbitrary amount rather than in either direction by a specific *amount*. A very simple application is that the occasionally useful, if extremely crude, truncation inequality $\mathbb{P}(X \geq \mathbb{E}(X))f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ for f positive, increasing and concave, complements Jensen's inequality $\mathbb{E}(f(X)) \leq f(\mathbb{E}(X))$; but may be extremely weak unless the probability of exceeding the expectation is controlled.

Measures of location and deviation probabilities tend to be monotone with respect to stochastic orders based on some form of domination and dispersion, respectively. Similarly probabilities of *exceeding* tend to be monotone with respect to stochastic orders based on some form of *skewness*. To control probabilities of exceeding, say, the expectation within some family of distributions one should therefore only need to consider maximal/minimal elements with respect to an appropriate order.

In practice, showing that two distributions are comparable with respect to skewness tends to involve proving convexity or concavity of somewhat cumbersome functions, often specified in terms of integrals and inverses lacking elementary closed forms. Sometimes such comparisons are greatly simplified by sign-change counting techniques. We will have a look at an example of this in the case of the family of Beta-distributions, which turn out to be ordered in a natural way. As time permits we will also consider some consequences for probabilities relating to binomials.

Optimal Estimation of Gaussian Mixtures via Denoised Method of Moments

Monday 22
14:00 - 16:00
Room 11

Yihong Wu[1] and Pengkun Yang

[1]Statistics and Data Science, Yale University, New Haven, United States
yihong.wu@yale.edu

The Method of Moments is one of the most widely used methods in statistics for parameter estimation, obtained by solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or non-uniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analysis of method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

In this talk I will present some recent results for estimating Gaussians location mixtures with known or unknown variance. To overcome the aforementioned theoretical and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space before executing the method of moments. Not only does this regularization ensures existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in Wasserstein distance via polynomial interpolation and marjorization, we establish the statistical guarantees and optimality of the proposed procedure. These results can also be viewed as provable algorithms for Generalized Method of Moments which involves non-convex optimization. Extensions to multiple dimensions will be discussed.

This is based on joint work with Pengkun Yang (Princeton).

Monday 22
14:00 - 16:00
Seminari B

Data-Driven Kaplan-Meier One-Sided Two-Sample Tests

Grzegorz Wyłupek[1]

[1]Institute of Mathematics, University of Wrocław, Wrocław, Poland
wylupek@math.uni.wroc.pl

The paper discusses existing approaches, known from the literature, to detection of stochastic ordering of the two survival curves as well as poses and solves the novel testing problem on it. Specifically, the null hypothesis asserts the lack of the ordering, while the alternative expresses its existence. An introduced test statistic is a functional of the standardized two-sample Kaplan-Meier process sampling in a randomly selected number of the random points being the observed survival times in the pooled sample and exploits the information contained in a specially defined one-sided weighted-log rank statistic. It automatically weights the magnitude and sign of their components becoming a sensible procedure in the considered testing problem. As a result, the corresponding test asymptoticly controls the errors of the both kinds at the specified significance level α. The conducted simulation study shows that the errors are also satisfactorily controlled when sample sizes are finite. Furthermore, in the comparison to the best and most popular tests, the new solution turns out to be a promising procedure which improves them upon. A real data analysis confirms that findings.

Monday 22
16:30 - 18:30
Room 6

Scan B-Statistic for Kernel Change-Point Detection

Yao Xie[1] and Shuang Li and Le Song and Hanjun Dai

[1]Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, United States
yao.c.xie@gmail.com

Detecting the emergence of an abrupt change-point is a classic problem in statistics and machine learning. Kernel-based nonparametric statistics have been used for this task which

enjoy fewer assumptions on the distributions than the parametric approach and can handle high-dimensional data. We focus on the scenario when the amount of background data is large, and propose two related computationally efficient kernel-based statistics for change-point detection, which are inspired by the recently developed B-statistics. A novel theoretical result of the paper is the characterization of the tail probability of these statistics using the change-of-measure technique, which focuses on characterizing the tail of the detection statistics rather than obtaining its asymptotic distribution under the null distribution. Such approximations are crucial to control the false alarm rate, which corresponds to the significance level in offline change-point detection and the average-run-length in online change-point detection. Our approximations are shown to be highly accurate. Thus, they provide a convenient way to find detection thresholds for both offline and online cases without the need to resort to the more expensive simulations or bootstrapping. We show that our methods perform well on both synthetic data and real data.

Applying Additive Hazards Model to Learn from Electronic Health Data

Monday 22
14:00 - 16:00
Room 7

Ronghui Xu[1] and Andrew Ying and Jue Hou and Jelena Bradic

[1]University of California, San Diego, La Jolla, CA, United States
rxu@ucsd.edu

Our work was motivated by the analysis project using the linked US SEER-Medicare database to studying treatment effects in men of age 65 years or older who were diagnosed with prostate cancer. Such data sets contain up to 100,000 human subjects and over 20,000 claim codes. The data are obviously not randomized with regard to the treatment of interest, for example, radical prostatectomy versus conservative treatment. Instrumental variable (IV) is an essential tool for addressing unmeasured confounding in this type of observational data. We develop an appropriate IV approach using the additive hazards model for patient survival, correcting the error in the literature using the Cox model as well as assuming a linear relationship between the binary treatment and the IV. Additionally, we have previously shown that the high dimensional claim codes contain rich information about the patients' non-cancer mortality, possibly capturing some of the unmeasured confounding beyond the usual clinical variables. We describe our approaches to making inference about the treatment effect using the high dimensional claim codes.

On Frequentist Coverage Errors of Bayesian Credible Sets in Moderately High Dimensions

Wednesday 24
10:30 - 12:30
Room 12

Keisuke Yano[1] and Kengo Kato

[1]Department of Statistics and Data Science, The University of Tokyo, Tokyo, Japan
kk976@cornell.edu

In this paper, we study frequentist coverage errors of Bayesian credible sets for an approximately linear regression model with (moderately) high dimensional regressors, where the

dimension of the regressors may increase with but is smaller than the sample size. Specifically, we consider Bayesian inference on the slope vector by fitting a Gaussian distribution on the error term and putting priors on the slope vector together with the error variance. The Gaussian specification on the error distribution may be incorrect, so that we work with quasi-likelihoods. Under this setup, we derive finite sample bounds on frequentist coverage errors of Bayesian credible rectangles. Derivation of those bounds builds on a novel Berry–Esseen type bound on quasi-posterior distributions and recent results on high-dimensional CLT on hyperrectangles. We use this general result to quantify coverage errors of Castillo–Nickl and L^{∞}-credible bands for Gaussian white noise models, linear inverse problems, and (possibly non-Gaussian) nonparametric regression models. In particular, we show that Bayesian credible bands for those nonparametric models have coverage errors decaying polynomially fast in the sample size, implying advantages of Bayesian credible bands over confidence bands based on extreme value theory.

Tuesday 23
14:00 - 16:00
Room 12

Estimation of Subgraph Densities in Noisy Networks

Qiwei Yao[1] and Jinyuan Chang and Eric Kolaczyk

[1]Statistics, London School of Economics, London, United Kingdom
q.yao@lse.ac.uk

While it is common practice in applied network analysis to report various standard network summary statistics, these numbers are rarely accompanied by some quantification of uncertainty. Yet any error inherent in the measurements underlying the construction of the network, or in the network construction procedure itself, necessarily must propagate to any summary statistics reported. Here we study the problem of estimating the density of an arbitrary subgraph, given a noisy version of some underlying network as data. Under a simple model of network error, we show that consistent estimation of such densities is impossible when the rates of error are unknown and only a single network is observed. Next, focusing first on the problem of estimating the density of edges from noisy networks, as a canonical prototype of the more general problem, we develop method-of-moment estimators of network edge density and error rates for the case where a minimal number of network replicates are available. These estimators are shown to be asymptotically normal as the number of vertices increases to infinity. We also provide confidence intervals for quantifying the uncertainty in these estimates based on either the asymptotic normality or a bootstrap scheme. We then present a generalization of these results to higher-order subgraph densities, and illustrate with the case of two-star and triangle densities. Bootstrap confidence intervals for those high-order densities are constructed based on a new algorithm for generating a graph with pre-determined counts for edges, two-stars, and triangles. The algorithm is based on the idea of edge-rewiring, and is of some independent interest. We illustrate the use of the proposed methods in the context of gene coexpression networks.

Develop a Risk Prediction Model for Depression Based on Text/Documents Mining Techniques

Thursday 25
14:00 - 16:00
Seminari B

Chun Yi Yeh[1] and Charlotte Wang

[1]Department of Mathematics, Tamkang University, New Taipei City, Taiwan
v777007@gmail.com

With the development of population, mental illness is getting more and more attention nowadays, especially the depression. According to WHO statistics, depression is affected by more than 300 million people of all age. Depression is on the rise globally and it will lead to suicide at its worst. Hence, how to prevent depression becomes an important issue in public health. In this study, we developed a risk prediction model for depression based on text mining and sentiment analysis. First, we utilized text mining techniques to structure documents. Second, a scoring method was proposed to identify, extract and quantify the emotion of documents through sentiments analysis. Final, we built up a risk prediction model for depression and evaluated the degree of the symptoms of depression for subjects. Eighty-seven lyrics were analyzed and were applied to demonstrate the proposed model. These lyrics were written by a Korean singer-songwriter who suffered from depression and committed suicide in 2017 after depression got worse. The proposed risk prediction model can evaluate people's emotion change through their diaries, essays, blogs, lyrics and so on, as well as, this model can detect their conditions of symptoms getting worse or getting better in advance. We think the model will be useful in public health for early detect and treatment and prevention on suicide.

Modeling and Analysis of Correlated Data using Pairwise Likelihood

Tuesday 23
10:30 - 12:30
Room 11

Grace Yi[1]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada
yyi@uwaterloo.ca

Correlated data arise commonly in practice, and modeling and analysis of such data have attracted extensive research interest. Although many methods have been developed, research gaps remain due to emerging issues. For instance, for correlated data with complex structures, modelling complexity is a serious issue, and it is desirable to develop flexible models that are both computationally manageable and interpretably meaningful. In terms of estimation, much research has been directed to estimation of the mean parameters with the association parameters treated as nuisance. There is relatively less work concerning both the marginal and association structures, especially in the semiparametric framework. In this talk, I will describe some methods of handling correlated data which are developed based on the pairwise likelihood formulation. Associated modeling strategies and estimation procedures will be discussed.

Thursday 25
14:00 - 16:00
Room 12

Exact MCMC with Differentially Private Moves

Sinan Yıldırım[1] and Beyza Ermiş

[1]Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey
sinanyildirim@sabanciuniv.edu

We view the penalty algorithm of Ceperley and Dewing (J Chem Phys 110(20):9812–9820, 1999), a Markov chain Monte Carlo algorithm for Bayesian inference, in the context of data privacy. Specifically, we studied differential privacy of the penalty algorithm and advocate its use for data privacy. The algorithm can be made differentially private while remaining exact in the sense that its target distribution is the true posterior distribution conditioned on the private data. We also show that in a model with independent observations the algorithm has desirable convergence and privacy properties that scale with data size. Two special cases are also investigated and privacy-preserving schemes are proposed for those cases: (i) Data are distributed among several users who are interested in the inference of a common parameter while preserving their data privacy. (ii) The data likelihood belongs to an exponential family. The results of our numerical experiments on the Beta-Bernoulli and the logistic regression models agree with the theoretical results.

Tuesday 23
16:30 - 17:30
Second Floor

Depth for Riemannian Manifold-Valued Functional Data Based on Optimal Transport

Ke Yu[1] and James Taylor

[1]University of Oxford, Oxford, United Kingdom
ke.yu.dphil@said.oxford.edu

Recently, interest has grown in functional data analysis for nonlinear Riemannian manifolds. An important example is the sphere-valued functional data, which naturally arises when data on a sphere has a time component. We propose notions of depth for Riemannian functional data which take into account the intrinsic geometry of data. The concept is based on three main building blocks. Firstly, we consider taking the Monge-Kantorovichs optimal transport in geodesic spaces with the reference distribution being a uniform distribution. Secondly, we rank the image of the optimal map on the sphere by depth for directional data. Thirdly, we work in the square-root velocity framework and solve the optimal transport problem in the tangent space TcM. We apply our proposal to classification, and compare its performance to classical approaches.

On the Robustness of Gradient-Based MCMC

Monday 22
10:30 - 12:30
Room 9

Samuel Livingstone and Giacomo Zanella[1]

[1]Department of Decision Sciences, Bocconi University, Milan, Italy
giacomo.zanella@unibocconi.it

We analyze the tension between robustness and efficiency for Markov chain Monte Carlo (MCMC) sampling algorithms. In particular, we focus on the robustness of MCMC algorithms with respect to heterogeneity in the target, an issue of great practical relevance but still understudied theoretically. We show that the spectral gap of the Markov chains induced by classical gradient-based MCMC schemes (e.g. Langevin and Hamiltonian Monte Carlo) goes to zero exponentially fast with the mismatch between the scale of the proposal and the one of the target, while for random walk Metropolis (RWM) the corresponding convergence to zero is linear. This result provides theoretical support to the well-known fact that gradient-based MCMC schemes, despite being more efficient than random walk ones, are also less robust to heterogeneity and more sensitive to hyper-parameters tuning (e.g. choice of preconditioning or mass matrix). Motivated by these considerations, we propose a novel and simple to implement gradient-based MCMC algorithm, inspired by the classical Barker accept-reject rule, with improved robustness properties. Extensive theoretical results, dealing with robustness to heterogeneity, geometric ergodicity and scaling with dimensionality, show that the novel scheme combines the robustness of RWM with the efficiency of classical gradient-based schemes. The theoretical results are illustrated with simulation studies.

Semiparametric Modeling of the Mean,Variance and Scale Parameters in Skew Normal Regression Models: A Bayesian Perspective

Monday 22
14:00 - 16:00
Room 10

Hector Zarate[1]

[1]Banco de la República, Bogota, Colombia
hzaratso@banrep.gov.co

The main goal of this paper is to integrate various statistical methods to estimate location, scale and shape functions in heteroscedastic semiparametric models when the response variable follows a skew normal distribution. We rely on the connection among smoothing methods, mixed models and a Bayesian Markov Chain sampling simulation methodology. The implications of our strategy lies in its potential to contribute to a simple and unified computational methodology that takes into account the factors that affect the parameters in the responses, which is crucial for a correct inference avoiding the requirement of fully parametric models. A simulation study investigates the performance of the estimates. Finally, an empirical application using forecasting predictive densities illustrates the benefits of our approach.

Friday 26
09:00 - 11:00
Room 5

Valid Properties of Truncated Student-t Distribution with Applications in the Analysis of Censored Data

Chi Zhang[1] and Guo-Liang Tian and Xiao Ke and Man-Lai Tang

[1]College of Economics, Shenzhen University, Shenzhen, China
chizhang@szu.edu.cn

Kim (2008) introduced an incorrect stochastic representation (SR) for the truncated Student-t (Tt) random variable. By pointing out that the gamma mixture based on a truncated normal distribution actually cannot result in a true Tt distribution, in this paper, we first propose three correct SRs and then recalculate the corresponding moments of the Tt distribution. Different from those derived from the invalid SR of Kim (2008), the correct moments of the Tt distribution play a crucial role in parameter estimations. Based on the proposed third SR and the correct expressions of truncated moments, expectation–maximization (EM) algorithms are developed for calculating the maximum likelihood estimates of parameters in the Tt distribution. Extensions to a Tt regression model and a t interval-censored regression model are provided as well. Simulated experiments are conducted to evaluate the performance of the proposed methods. Finally, two real data analyses corroborate the theoretical results.

Wednesday 24
10:30 - 12:30
Room 10

Asymptotic Independence of Spiked Eigenvalues and Linear Spectral Statistics for Large Sample Covariance Matrices

Zhixiang Zhang[1] and Guangming Pan

[1]MAS/SPMS, Nanyang Technological University, Singapore
S160011@e.ntu.edu.sg

We consider a general spiked sample covariance model and prove a central limit theorem of spiked eigenvalues under the regime that the dimension and sample size grow to infinity proportionally.This result is an extension of previous central limit theorem established for spiked eigenvalues under the assumption that the population covariance matrix has a block diagonal structure.The proof rests on a central limit theorem for a random quadratic form, which can be of independent interest.Then the relationship of spiked eigenvalues and linear spectral statistics are considered. We show that they are asymptotically independent. Some corrected consistent estimators for spiked population eigenvectors are also proposed.

Second Order Stein: SURE for SURE and Other Applications

Monday 22
14:00 - 16:00
Room 11

Cun-Hui Zhang[1] and Pierre Bellec

[1]Department of Statistics, Rutgers University, USA
czhang@stat.rutgers.edu

Stein's formula states that a random variable of the form $z^\top f(z) - \text{div} f(z)$ is mean-zero for all functions f with integrable gradient. Here, $\text{div} f$ is the divergence of the function f and z is a standard normal vector. We develop Second Order Stein formulas for statistical inference with high-dimensional data. In the simplest form, the Second Order Stein formula characterizes the variance of $z^\top f(z) - \text{div} f(z)$. A first application of the Second Order Stein formula is an Unbiased Risk Estimate for Stein's Unbiased Risk Estimator (SURE for SURE): an unbiased estimate provides information about the squared distance between SURE and the prediction error in the Gaussian sequence model. SURE for SURE has a simple form and can be computed explicitly for almost differentiable estimators, for example the Lasso and the Elastic Net. Other applications of the Second Order Stein formula are provided in high-dimensional regression. This includes novel bounds on the variance of the size of the model selected by the Lasso, and a general semi-parametric scheme to de-bias an almost differentiable initial estimator in order to estimate a low-dimensional projection of the unknown regression coefficient vector.

Weight Choice for Composite and Model-Averaged Estimation

Monday 22
10:30 - 12:30
Room 8

Jing Zhou[1] and Gerda Claeskens

[1]ORStat, KU Leuven, Leuven, Belgium
jing.zhou@kuleuven.be

A composite estimator is constructed by weighting multiple loss functions whereas a model-averaged estimator is obtained by weighting multiple estimators. The efficiency of the estimators is affected by the weight choice. We consider the approximate message passing algorithm (Bradic, 2016; Bayati and Montanari, 2011; Donoho and Montanari, 2016) which gives the expression of the asymptotic mean square error (AMSE) of the estimators when both the sample size n and the dimension p grow to infinity with the ratio $n/p \to \delta$.

We mainly consider the sparse high-dimensional linear model setting where $\delta \in (0, 1)$ with the l_1-regularized composite and model-averaged quantile estimator. We focus on finding the selection-incorporated weights by minimizing the AMSEs without assuming a perfect selection of the active set of coefficients. Our main contributions focus on the l_1-regularized model-averaged estimator where the model-averaged quantile estimator is a special case. We obtain the expression of the AMSE of the l_1-regularized model-averaged estimator by taking the correlation between the estimators into account. The analytical expression of the selection-incorporated weight is obtained by minimizing the expression of the AMSE as a quadratic function of the weight. Additionally, we derive a Stein-type estimator of the AMSE for the model-averaged estimator which can be used in practice.

For a dense linear model with $\delta \in [1, \infty)$, following Donoho and Montanari (2016) and Bayati and Montanari (2011), the asymptotic variance of the model-averaged estimator coincides with the AMSE. We show that it shares a similar structure with the AMSE in the high-dimensional case and can be estimated in practice. Similar to Donoho and Montanari (2016), we identify that the asymptotic variance of the model-averaged estimator also contains extra Gaussian noise components which are due to the high dimensionality.

Author Index

UNIVERSITÀ DEGLI STUDI DI PALERMO

dSEAS dipartimento scienze economiche aziendali e statistiche department of economics business and statistics

POLITECNICO DI TORINO

SIS Società Italiana di Statistica

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Università degli Studi di Cagliari

Bernoulli Society for Mathematical Statistics and Probability

Springer

CAMBRIDGE UNIVERSITY PRESS