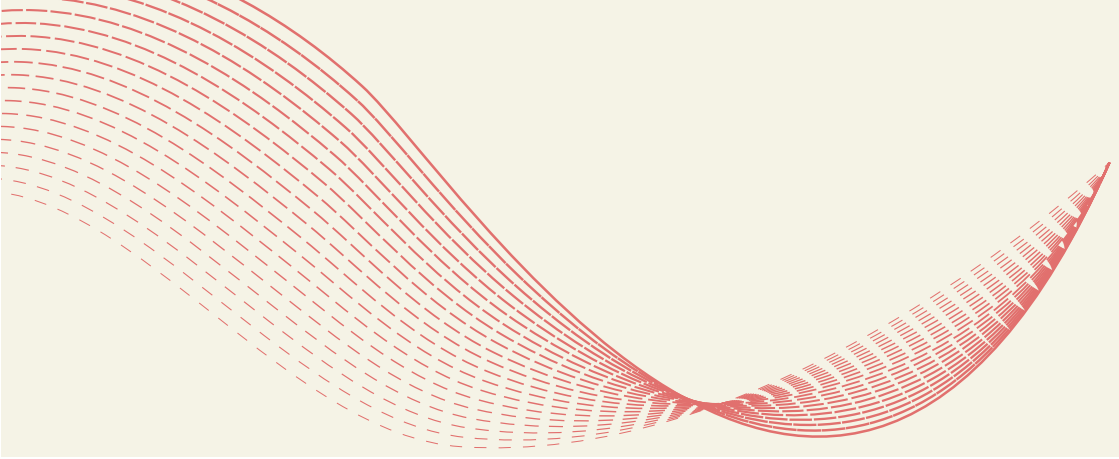
ISNPS2018

4th Conference of the International Society for Nonparametric Statistics

June 11–15 2018
Salerno, Italy



Title: Book of Abstracts - ISNPS 2018
Abstracts of the the 4th Conference of the International Society for Nonparametric Statistics, held in Salerno (Italy) 11-15 June 2018 (`http://www.isnps2018.it`)

Edited by: Michele La Rocca, Brunero Liseo, Maria Lucia Parrella, Luigi Salmaso, Luca Tardella

Year: 2018
ISBN: 978-88-61970-00-7

ISNPS2018

4th Conference of the International Society for Nonparametric Statistics

June 11–15 2018
Salerno, Italy

ISNPS 2018 Conference Co-chairs

Michele La Rocca
Brunero Liseo
Luigi Salmaso

Scientific Committee

Patrice Bertail
Efstathia Bura
Walter Distaso
Bing Li
Yanyuan Ma
Marianna Pensky
Michael G. Schimek
Piercesare Secchi
Stefan Sperlich

Local Organizing Committee

Alessandra Amendola
Rosa Arboretti
Stefano Bonnini
Eleonora Carrozzo
Livio Corain
Pietro Coretto
Francesco Giordano
Marcella Niglio
Maria Lucia Parrella
Cira Perna
Maria Rizzo
Luca Tardella

FOREWORDS FROM THE CO-CHAIRS OF THE 2018 ISNPS CONFERENCE

Following the successful ISNPS Conferences in Chalkidiki (Greece), Cadix (Spain) and Avignon (France) we are pleased to welcome you to the 4th ISNPS Conference. The meeting is co-sponsored by the Institute of Mathematical Statistics (IMS), the Bernoulli Society for Mathematical Statistics and Probability, the Nonparametric Statistics Section of the American Statistical Association, Cytel, the Italian Statistical Society (SIS), the University of Padua, Sapienza University of Rome, and the University of Salerno.

The conference is a collective effort and many individuals and organizations contributed. The Scientific Committee, the organizers of the Invited sessions, the Local organizing committee, the local hosting universities and volunteers, all of them played a substantial role in the organization of the conference. We warmly acknowledge their work and the support of all the institutions and organizations that have sponsored the conference.

The conference is hosted by the city of Salerno, in Italy. The venue has a strategic location which allows participants to enjoy the natural and cultural attractions of the city and its surroundings. Participants can quickly reach and visit the worldwide known sites of the Campania Region, like the Amalfi Coast (with the beautiful villages of Positano and Ravello), the Cilento National Park (both of them being part of the UNESCO heritage), and the fantastic archaeological areas of Pompei, Paestum and Herculaneum.

As in the previous editions, the 4th ISNPS conference will put together recent advances and trends in several areas of nonparametric statistics. With 12 plenary and special invited sessions, 69 invited sessions, 30 contributed sessions, with about 450 participants from all over the world, ISNPS 2018 will be the perfect place to facilitate the exchange of research ideas, promote collaboration between researchers and contribute to the further development of the field. The scientific program, scheduled from June 11 to the June 15 with a pause in the afternoon of the 13th, will include plenary and special invited talks, invited talks as well as contributed talks and posters on all areas of nonparametric statistics. Included topics are Hypothesis testing, Statistical learning, Big data, Adaptive inference, Sparsity, High-dimensional statistical inference and tests, Nonparametric time-series, Object Data Analysis, Extreme values and Risk, Complex data, Functional analysis, Density estimation, Statistical inverse problems, Model selection and validation, Survival analysis, Permutation and Resampling techniques, Applications of nonparametric statistics.

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

Michele La Rocca, Brunero Liseo and Luigi Salmaso

PROGRAM SCHEDULE

From	to	Monday, June 11	Tuesday, June 12	Wednesday, June 13	Thursday, June 14	Friday, June 15
8:30	9:00	Registration and introduction				
9:00	9:30	Plenary session	Invited parallel sessions	Invited parallel sessions	Invited parallel sessions	Invited parallel sessions
9:30	10:00					
10:00	10:30	Special invited session / Contributed sessions				
10:30	11:00		Coffee break	Coffee break	Coffee break	Coffee break
11:00	11:30	Coffee break	Plenary session	Special invited session / Contributed sessions	Plenary session	Plenary session
11:30	12:00	Invited parallel sessions				
12:00	12:30		Special invited session / Contributed sessions	Invited parallel sessions	Special invited session / Contributed sessions	Special invited session / Contributed sessions
12:30	13:00					
13:00	13:30					
13:30	14:00	Lunch	Lunch	Optional Excursions	Lunch	Lunch
14:00	14:30					
14:30	15:00		Invited parallel sessions		Invited parallel sessions	Invited parallel sessions
15:00	15:30	Special invited session / Contributed sessions				
15:30	16:00					
16:00	16:30	Coffee break				
16:30	17:00	Invited parallel sessions	Coffee break		Coffee break	
17:00	17:30		Special invited session / Contributed sessions		Special invited session / Contributed sessions	Closing and Farewell party
17:30	18:00					
18:00	18:30				ISNPS MEETING	
18:30	19:00	Poster session	Invited parallel sessions			
19:00	19:30					
20:00		Welcome reception			Social Dinner	

GENERAL INFORMATION

Conference venue

The 4th ISNPS conference will take place at the Grand Hotel Salerno. Plenary sessions and Special invited sessions will take place in the Tafuri Auditorium, located on the ground floor. Parallel sessions will take place on the first floor in the Vietri, Ravello, Furore, Positano, Amalfi, Maiori and Procida Rooms. Posters will be located in the Foyer Minori, in front of the Amalfi Room. Detailed information on the invited sessions and invited speakers, as well as the full program with abstracts, is available on the website of the conference `http://www.isnps2018.it/`. A dedicated app (ISNPS2018) for the meeting can be downloaded from the Apple Store (for iOS) and Play Store (for Android).

Registration

The ISNPS staff will be at the registration desk in the Atena Room on the ground floor of the Grand Hotel Salerno. Registration will start at 6:00 pm on Sunday June 10 and will be open for the whole time of the conference.

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 (Acrobat) or PPT (PowerPoint) format on a USB memory stick. Speakers in invited sessions have 30 minutes, a maximum of 25 minutes for their presentation and 5 minutes for discussions. Speakers in contributed sessions have 20 minutes, 15 minutes for their presentation and 5 minutes for discussions. Chairs are kindly requested to keep strictly to the schedule. Papers should be presented in the order they are listed in the program for the convenience of attendees, who may wish to go to the other rooms mid-session to hear particular speakers talks. In case of absence of the presenter, please use the extra-time for a break or a discussion, in order to remain on schedule. IT technicians from the conference organization will be available during the conference. Please refer to the registration desk in case of problems.

Social events

The fees cover the Coffee breaks, the Welcome Reception and the Farewell Party. The Welcome Reception is scheduled on Monday, June 11, at 8:00 pm while the Farewell Party is scheduled on Friday, June 15 at 4:30 pm. They will take place on the Roof Terrace of the Grand Hotel Salerno.
The Conference Dinner will take place at the Hotel Raito in Raito (a charming village on the Amalfi coast) on Thursday June 14 at 8:00 pm. Please note that

it is not included in the conference fee. Please check your conference booking confirmation to find out if you have registered to attend the dinner. The Conference Dinner can also be booked on site from Monday June 11 to Tuesday June 12 at the conference registration desk, at the price of € 70.00. The meeting point for transfer to Raito by private bus will be at 8:00 pm at the Grand Hotel Salerno. Staff will be collecting dinner tickets on arrival at the Hotel Raito. The dinner is expected to end around midnight. Special dietary requirements, vegetarian and vegan meals must be indicated at the registration desk in advance.

Optional excursions

The afternoon of Wednesday June 13 is free. A local travel agency, Barbirotti Viaggi offers three types of excursions on request: to the archaeological sites of Pompei and Paestum and to the village of Ravello on the Amalfi coast. Detailed information and prices are available at `http://www.isnps2018.it/#events`
For information and reservations, please contact:
Barbirotti Viaggi
E-mail: prenotazioni@barbirottiviaggi.it
Office +39 089 99 53 237
Mobile +39 339 52 02 517
`www.barbirottiviaggi.it`

Accommodation

Accommodation is not included in your conference registration.

Meals and refreshments

Tea and Coffee breaks will be served in the conference area according to the program. Luncheons will be served at the restaurant Sala Ninfea on the first floor of the Grand Hotel Salerno, just in front of the conference rooms. If you have not reserved meals at registration time, you can still buy your ticket meals at the hotel reception one day in advance. We kindly ask you to wear your conference badge during the meeting and to show your lunch ticket to the restaurant staff.

Wireless internet connection

Wireless internet is available across the conference area, enabling participants to use their laptops to connect to the internet. The name of the network and the password will be available on site.

EXIBITORS

Springer

Springer is a leading global scientific, technical and medical portfolio, providing researchers in academia, scientific institutions and corporate R&D departments with quality content through innovative information, products and services.

Cytel

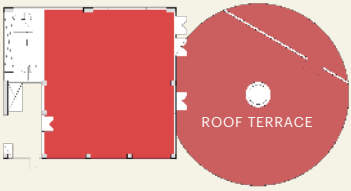
Cytel is shaping the future of drug development. As the world's largest independent clinical biostatistics research and development organization, Cytel helps leading pharmaceutical, biotech and medical device companies improve clinical success rates via optimal study design, effective data management, and accurate statistical analysis. Cytel provides both software solutions for the design and analysis of clinical trials, including industry standards East®, StatXact® and LogXact®, as well as data focused clinical research services. With operations across North America, Europe, and India, Cytel employs 900 professionals, with strong talent in biostatistics, programming, and data management.

PROCEEDINGS

The 4^{th} ISNPS Conference Chairs have agreed with Springer the publication of a volume with the Proceedings of the Fourth ISNPS Conference. Instructions and deadlines will be communicated on the Conference website and sent by email to all registered participants.

MAP OF CONFERENCE ROOMS

ROOF TERRACE



FIRST FLOOR

AUDITORIUM TAFURI

GROUND FLOOR

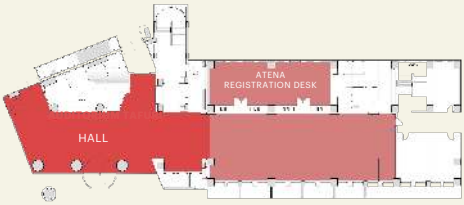


Table of contents

DETAILED PROGRAM

Plenary Session PL-01 • Monday 11, 09:00-10:00 • Auditorium Tafuri

WITH GREAT POWER COMES GREAT RESPONSIBILITY: MULTIVARIATE PERMUTATION TESTS AND THEIR NUMERICAL IMPLEMENTATION

by Fortunato Pesarin[1] & Philip B. Stark[2]
[1]University of Padua, Italy; [2]University of California, Berkeley, USA

Chair: Luigi Salmaso (University of Padua)

Abstract: Nonparametric combination (NPC) of dependent tests, combined with the Union-Itersection (UI) principle of S.N. Roy (1953), allows multivariate tests to be derived in great generality. UI assumes that the hypotheses H_0 and H_1 can be equivalently written "componentwise" as $H_0 \equiv \bigcap_{k=1}^{K} H_{0k}$ and $H_1 \equiv \bigcup_{k=1}^{K} H_{1k}$.
A global test derives from suitable partial tests $\{T_k\}$ of H_{0k} versus $H_{1k}; k = 1, \ldots, K$.
We discuss two classic problems to illustrate the theory.
The K partial tests are generally dependent in a way that is difficult or impossible to characterize in closed form, but NPC can be calibrated by simulation using pseudo-random permutations "synchronized" across dimensions.
To keep the advantages that nonparametric tests afford, the simulation needs to be adequately accurate – which is not merely a question of the number of replications. Surprisingly, default pseudo-random number generators (PRNGs) in common statistical packages are generally inadequate for modern permutation inference, and sampling algorithms based on those PRNGs vary widely in efficiency and accuracy. The default PRNGs in statistical packages should be replaced with cryptographically secure PRNGs (CS-PRNGs), and in some cases, the algorithms for generating random integers and random samples should be replaced with better methods.

Contributed Session CS-01 • Monday 11, 10:00-11:00 • Room Ravello

RANK DATA

Chair: Abu Minhajuddin (University of Texas Southwestern)

Title: ***Inference for covariate-adjusted semiparametric Gaussian copula model using residual ranks***
Speaker: **Yue Zhao**, KU Leuven, Belgium
Co-author(s): Irene Gijbels, Ingrid Van Keilegom
Schedule: *Monday 11 June, 10:00–10:20*
Abstract: We investigate the inference of the copula parameter in the semiparametric Gaussian copula model when the copula component, subject to the influence of a covariate, is only indirectly observed as the response in a linear regression model. We consider estimators based on residual ranks instead of the usual but unobservable oracle ranks. We first study two such estimators for the copula correlation matrix, one via inversion of Spearman's rho and the other via normal scores rank correlation coefficient. We show that these estimators are asymptotically equivalent to their counterparts based on the oracle ranks. Then, for copula correlation matrix under constrained parametrizations, we show that the classical one-step estimator in conjunction with the residual ranks remains semiparametrically efficient for estimating the copula parameter. The accuracy of the estimators based on residual ranks is confirmed by simulation studies.

Title: ***Semiparametric multi-parameter regression survival modelling***
Speaker: **Frank Eriksson**, University of Copenhagen, Denmark
Co-author(s): Kevin Burke, Christian Pipper
Schedule: *Monday 11 June, 10:20–10:40*
Abstract: We consider a log-linear model for survival data, where both the location and scale parameters depend on covariates and the baseline hazard function is completely unspecified. This model provides the flexibility needed to capture many interesting features of survival data at a relatively low cost in model complexity. Estimation procedures are developed and asymptotic properties of the resulting estimators are derived using empirical process theory. Finally, a resampling procedure is developed to estimate the limiting variances of the estimators. The finite sample properties of the estimators are investigated by way of a simulation study, and a practical application is briefly illustrated.

Title: ***When Wilcoxon-Mann-Whitney Fails***

Speaker: **Abu Minhajuddin**, University of Texas Southwestern, United States of America
Schedule: *Monday 11 June, 10:40–11:00*
Abstract: A commonly used alternative to Student's two independent sample t-test is the Wilcoxon-Mann-Whitney U (WMW) test. Often time, when the assumption of normality is not tenable, researchers use WMW as an alternative analysis. This rank-based, easy to implement test has a lot of attractive attributes. But this should not be the best alternative test to use anytime normality is violated. In this presentation, we will show simulation based evidence to situations when WMW is not the more powerful test of choice. We'll also recommend a simple procedure for users to decide when to use WMW in practice and when not to use. Additionally, we'll suggest some alternative tests to use when WMW fails.

Contributed Session CS-02 • Monday 11, 10:00-11:00 • Room Furore

PERMUTATION TESTS FOR EXPERIMENTAL AND OBSERVATIONAL STUDIES
Chair: Olivier Renaud (University of Geneva)

Title: ***Identifying physiological correlates of emotional response in children: a nonparametric combination based approach in a multiple-baseline design***
Speaker: **Chiara Brombin**, Vita-Salute San Raffaele University, Milano, Italy
Co-author(s): Simona Scaini, Clelia Di Serio
Schedule: *Monday 11 June, 10:00–10:20*
Abstract: This contribution presents results from an experimental study carried out in a sample of children, recruited from the general population, and aimed at identifying physiological correlates of emotions induced by affective visual stimuli. Twelve stimuli (4 neutral pictures and 8 emotionally charged pictures) were randomly administered to each participant. During stimuli presentation, both subjective responses (e.g., the felt emotion and its intensity, as well as ratings on pleasure, arousal, and dominance dimensions) and objective physiological measurements were collected. The goal of the study is twofold. First, the study aims at evaluating whether neutral stimuli elicit comparable physiological responses, thus representing a homogeneous emotional category. Second, it aims at assessing whether responses to emotionally charged pictures are significantly different from those elicited by neutral pictures. The nonparametric

combination (NPC) methodology proposed by Pesarin and Salmaso (2010) has been applied to effectively deal with repeated measures and generalized to account for multiple baseline conditions, since 4 neutral stimuli, representing our reference condition, are presented. Within this framework, the effectiveness of stimuli may be tested at different levels, involving sensible subgroups of variables or pictures. The flexible approach here applied may provide new insights into research on emotions involving (visual) stimulation.

Title: ***Multiplication-Combination Tests for Incomplete Paired Data***
Speaker: **Lubna Amro**, Ulm University, Germany
Co-author(s): Frank Konietschke, Markus Pauly
Schedule: *Monday 11 June, 10:20–10:40*
Abstract: We consider statistical procedures for hypothesis testing of real valued functionals of matched pairs with missing values. In order to improve the accuracy of existing methods, we propose a novel multiplication combination procedure. Dividing the observed data into dependent (completely observed) pairs and independent (incompletely observed) components, it is based on combining separate results of adequate tests for the two sub datasets. Our methods can be applied for parametric as well as semi- and nonparametric models and make efficient use of all available data. In particular, the approaches are flexible and can be used to test different hypotheses in various models of interest. This is exemplified by a detailed study of mean- as well as rank-based approaches. Extensive simulations show that the proposed procedures are more accurate than existing competitors. A real data set illustrates the application of the methods.

Title: ***Permutation tests for the comparison of signals stemming from experimental designs***
Speaker: **Olivier Renaud**, University of Geneva, Switzerland
Co-author(s): Jaromil Frossard
Schedule: *Monday 11 June, 10:40–11:00*
Abstract: In many applications, researchers are interested in comparing signals that come from different experimental conditions, different groups or different settings. Moreover, they often want a precise assessment that stipulates for each time point and each (main or interaction) effect separately whether the difference is significant or not. We will present several permutation schemes that test any factor or covariate for each time point. We show that they can be included in many multiple comparison methods from the very basic Bonferroni to the cluster mass and the threshold-free cluster enhancement methods, which

control the FWER while holding a high statistical power. We also present an extension to models that include a random effect, e.g. of the subject, as well as its interactions with fixed effects and an extension to cross random effects, e.g. subjects and items in a psychology or neuroscience experiment. An R package including these methods is available on CRAN (`https://CRAN.R-project.org/package=permuco`)

Contributed Session CS-03 • Monday 11, 10:00-11:00 • Room Positano

SURVIVAL ANALYSIS

Chair: Miguel A. Delgado (Universidad Carlos III de Madrid)

Title: ***Cox model with nonparametric frailty for modeling hierarchical survival data***

Speaker: **Francesca Gasperoni**, MOX - Modelling and Scientific Computing Department of Mathematics, Politecnico di Milano, Milan, Italy

Co-author(s): Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples

Schedule: *Monday 11 June, 10:00–10:20*

Abstract: Clinical records are typical examples of hierarchical survival data, in which patients are statistical units grouped in hospitals or in cities. Semiparametric Cox models with a parametric shared frailty term are commonly used in these cases. The novel contribution of this work consists in a survival model with a nonparametric shared frailty term, which means a discrete random effect with an unknown number of elements in its support. Our aim is two-fold: on one hand, we want to propose a more flexible model for grouped survival data, on the other hand, we want to detect clusters of groups (i.e., clusters of hospitals), named latent populations, and characterize them a posteriori. Then, we propose a tailored Expectation-Maximization algorithm to estimate the number of latent populations, the frailty distribution, the proportion associated to each latent population and the classical parameters of a Cox model. To conclude, we show an application to a clinical administrative database, in which several information of patients suffering from Heart Failure is collected (like age, comorbidities, procedures etc). Applying our model, we are able to detect a latent clustering structure among hospitals and this result can be exploited by both clinicians and healthcare managers.

Title: ***Nonparametric estimation of the joint distribution of two gap times under various types of censoring and truncation***

Speaker: **Carla Maria Moreira**, EPIUnit - Institute of Public Health, University of Porto (ISPUP), Portugal
Co-author(s): Jacobo de Uña-Álvarez, Ana Cristina Santos
Schedule: *Monday 11 June, 10:20–10:40*
Abstract: In Survival Analysis and Epidemiology, among other fields, one typical goal is to estimate a joint distribution of successive times which may suffer from various types of censoring and truncation. Particular difficulties appear, for example, when information on a cohort is obtained through intermittent visits or successive cross-sections; in such cases, special combinations of left-truncated, right-censored and interval censored data will appear. We describe one of such complicated settings, and we introduce an inverse-probability-weighted type estimator (IPWE) for the joint distribution of two gap times which takes the aforementioned censoring and truncation issues into account. The performance of the proposed estimator is investigated through simulations. We also discuss asymptotic results. For illustration purposes, the estimator is applied to data from the EPIPorto adults cohort study.

Title: ***Standardization & decomposition in survival analysis***
Speaker: **Miguel A. Delgado**, Universidad Carlos III de Madrid, Spain
Co-author(s): Adrés García-Souza
Schedule: *Monday 11 June, 10:40–11:00*
Abstract: We provide standardization techniques for survival time distributions using right censored data that can be used to decompose the difference between distributional characteristics of two populations into structural and composition counterfactual effects. The structural effect is the difference that would had been observed if the relevant characteristics explaining survival time in the two populations had been identically distributed. The composition effect is the difference that would had been observed if the populations had only differed in the distribution of relevant characteristics. We propose two alternative standardization methods. One assumes a semiparametric proportional hazard specification. The other is based on a non-parametric specification of the joint distribution of lifetime and components, which is estimated using Kaplan-Meier weights. We pay particular attention to standardization and decomposition of the mean survival time and the hazard rate. Finite sample properties of the alternative methods are studied by means of Monte Carlo experiments. We apply the proposed techniques to study gender unemployment duration gaps using data from Spain.

Contributed Session CS-04 • Monday 11, 10:00-11:00 • Room Procida

NONPARAMETRIC INFERENCE AND ESTIMATION I

Chair: Radu V. Craiu (University of Toronto)

Title: ***A two-sample test for the equality of distributions with high-dimensional data***

Speaker: **Marta Cousido Rocha**, University of Vigo, Spain

Co-author(s): Jacobo de Uña Álvarez, Jeffrey D Hart

Schedule: *Monday 11 June, 10:00–10:20*

Abstract: A recurring theme in modern statistics is dealing with high-dimensional data whose main feature is that the data dimension p (number of variables) is large while the sample size is small. In this context our aim is to address the problem of testing the null hypothesis that the marginal distributions of each of the p variables is the same for two groups of individuals. We propose a two-sample test statistic which measures the overall distance between the empirical characteristic functions pertaining to the two groups along the p-variables. The asymptotic normality of the test statistic is derived under mixing conditions. In our asymptotic analysis the number of variables tends to infinity, while the sample sizes remain fixed. In order to apply the test in practice several estimators of the variance are proposed leading to three slightly different versions of the test. An alternative test based on the p-values computed from a permutation approach is also proposed. A simulation study to investigate the finite sample properties of the proposed tests is carried out. A practical illustration involving microarray data is provided.

Title: ***Block bootstrap for abrupt change in mean avoiding variance estimation***

Speaker: **Michal Pesta**, Charles University, Czech Republic

Co-author(s): Barbora Pestova

Schedule: *Monday 11 June, 10:20–10:40*

Abstract: We deal with sequences of weakly dependent observations that are naturally ordered in time. Their constant mean is possibly subject to change at most once at some unknown time point. The aim is to test whether such an unknown change has occurred or not. The change point methods presented here rely on ratio type test statistics based on maxima of the cumulative sums. These detection procedures for the abrupt change in mean are also robustified by considering a general score function. The main advantage of the proposed approach

is that the variance of the observations neither has to be known nor estimated. The asymptotic distribution of the test statistic under the no change null hypothesis is derived. Moreover, we prove the consistency of the test under the alternatives. A block bootstrap method is developed in order to obtain better approximations for the test's critical values. The validity of the bootstrap algorithm is shown. The results are illustrated through a simulation study, which demonstrates computational efficiency of the procedures. A practical application to real data is presented as well.

Title: ***Bayesian Inference for Conditional Copulas using Gaussian Process Single Index Models***
Speaker: **Radu V. Craiu**, University of Toronto, Canada
Co-author(s): Evgeny Levi
Schedule: *Monday 11 June, 10:40–11:00*
Abstract: Parametric conditional copula models allow the copula parameters to vary with a set of covariates according to an unknown calibration function. Flexible Bayesian inference for the calibration function of a bivariate conditional copula is introduced. The prior distribution over the set of smooth calibration functions is built using a sparse Gaussian process (GP) prior for the single index model (SIM). The estimation of parameters from the marginal distributions and the calibration function is done jointly via Markov Chain Monte Carlo sampling from the full posterior distribution. A new Conditional Cross Validated Pseudo-Marginal (CCVML) criterion is used to perform copula selection and is modified using a permutation-based procedure to assess data support for the simplifying assumption.

Special Invited Session SIS-01 • Monday 11, 10:00-11:00 • Auditorium Tafuri

ON A THEORY OF DISTRIBUTION-FREE TESTS: A NEW APPROACH

by Estate Khmaladze
University of Wellington, New Zealand

Speaker: Estate Khmaladze
Chair: Yanyuan Ma (Penn State University)

Abstract: The aim of the talk is to demonstrate how an application of unitary operators to function-parametric empirical processes leads to the new point of view on the theory of distribution-free tests of statistical hypotheses.

In particular, we will show how an empirical process in Rˆd can be transformed into uniform empirical process on $[0,1]^d$, both for continuous and discrete distributions; or how an empirical process in testing parametric families of distributions, which also may depend on random covariates, can be transformed into an empirical process with a standard asymptotic distribution, free from the covariates.
In the next example, we will consider the problems of parametric regression, where the regression empirical process again will be "rotated" by the corresponding unitary operator into an empirical process with asymptotic distributions, independent from behaviour of covariates.
In the last example, we will see the first possibilities to construct distribution-free tests for hypotheses about transition matrices of Markov chains. In particular, we will realise that, say, testing that a sequence is a Markov chain with finite number of states and positive transition matrix, i.e. such that p(y|x)>0 for all x and y, can be transformed to, and therefore is equivalent to testing that it is an i.i.d. sequence of random variables uniformly distributed on the set of states.

Invited Session IS-01 • Monday 11, 11:30-13:30 • Room Vietri

SURVIVAL ANALYSIS

Organizer(s): Ingrid Van Keilegom (KU Leuven)
Chair: Ingrid Van Keilegom (KU Leuven)

Title: ***Estimation of a bivariate conditional copula when a variable is subject to random right censoring.***
Speaker: **Anouar El Ghouch**, UCL, Belgium
Co-author(s): Taoufik Bouezmarni, Félix Camirand Lemyre
Schedule: *Monday 11 June, 11:30–12:00*
Abstract: This paper concerns the dependence structure of a random pair (Y, Z) conditionally upon a covariate X in a case where the variable Y is subject to random right censoring. The dependence structure is described by a conditional copula. We propose a nonparametric procedure to estimate this copula and we establish its asymptotic properties. The finite sample behavior of the proposed estimator is investigated in a numerical study. The methodology is also illustrated through a real data example featuring patients with malignant melanoma.

Title: ***Non-parametric cure rate estimation under insufficient follow-up using extremes***

Speaker: **Mikael Escobar-Bach**, KULeuven, Belgium

Co-author(s): Ingrid Van Keilegom

Schedule: *Monday 11 June, 12:00–12:30*

Abstract: Mixture cure models aim to represent survival data that might assume a sizable fraction of cured individuals with no expression of the event of interest. The cure fraction or cure rate represents a key value to estimate and has recently brought many attentions in several areas. However, most of the estimation methods proposed do not handle the case of insufficient follow-up, that is when the right-end point of the censors distribution is strictly less than the susceptibles one, and consequently lead to an over-estimation of the cure rate. In this talk, we intend to partially fill this gap by presenting a new estimator for the cure rate by use of extrapolation techniques from the Extreme Value Theory. We will establish its asymptotic properties and support its computational efficiency in a simulation study. A practical application is considered through a survival analysis of breast cancer data.

Title: ***The maximum likelihood estimator in the current status model***

Speaker: **Kim Hendrickx**, Hasselt University, Belgium

Co-author(s): Piet Groeneboom

Schedule: *Monday 11 June, 12:30–13:00*

Abstract: Survival models are commonly used to characterize the distribution of a variable Y that is not observed directly. Depending on what information is obtained on Y, different censoring schemes arise. In the current status model, the variable Y of interest is only known to lie before or after some random censoring variable T. Each observed sample consists of a set of n inspection times T_i and n censoring indicators $\Delta_i = 1_{\{Y_i \leq T_i\}}$. One could say that the ith observation indicates the *current status* of component i at time T_i. The maximum likelihood estimator (MLE) of the distribution function of Y converges at cube-root n-rate to Chernoff's distribution. It is known that the non-parametric bootstrap cannot be used to generate the asymptotic limiting distribution of the MLE under current status data. We propose to estimate the distribution function by a smoothed version of the MLE and discuss new ways of constructing pointwise bootstrap confidence intervals for the distribution function based on this smoothed MLE. We also elucidate other statistical applications under the current status model where the MLE is involved.

Title: ***On an extension of the promotion time cure model***

Speaker: **François Portier**, Télécom ParisTech, France
Co-author(s): Ingrid Van Keilegom, Anouar El Ghouch
Schedule: *Monday 11 June, 13:00–13:30*
Abstract: We consider the problem of estimating the distribution of time-to-event data that are subject to censoring and for which the event of interest might never occur, i.e., some subjects are cured. To model this kind of data in the presence of covariates, one of the leading semiparametric models is the promotion time cure model (Yakovlev (1996)), which adapts the Cox model to the presence of cured subjects. Estimating the conditional distribution results in a complicated constrained optimization problem, and inference is difficult as no closed-formula for the variance is available. We propose a new model, inspired by the Cox model, that leads to a simple estimation procedure and that presents a closed formula for the variance. We derive some asymptotic properties of the estimators and we show the practical behavior of our procedure by means of simulations. We also apply our model and estimation method to a breast cancer data set.

Invited Session IS-02 • Monday 11, 11:30-13:30 • Room Ravello

NEW DIRECTIONS IN SURVIVAL ANALYSIS AND SPATIAL INFERENCE
Organizer(s): Sophie Dabo-Niang & Akim Adekpedjou (University of Lille & University of Missouri)
Chair: Akim Adekpedjou (University of Missouri)

Title: ***Density estimation over spatio-temporal data streams***
Speaker: **Aboubacar Amiri**, Lille 3, France
Co-author(s): Sophie Dabo
Schedule: *Monday 11 June, 11:30–12:00*
Abstract: In the last few years, data can be collected extremely easily in many scientific research fields. This became possible by the recent technological advances that have made online monitoring possible. In such situations, if real time or online estimations are expected, the usual nonparametric techniques rapidly require a lot of time to be computed and therefore become useless in practice. Adaptative counterparts of the classical kernel density estima- tors, that can be updated extremely easily when a new set of observations is available are investigated, for spatio-temporal processes with weak dependence structures. Mean square, uniform almost sure convergences and a

central limit result are obtained under general and easily verifiable conditions. The efficiency of the considered estimators is eval- uated through simulations and a real data application. The results show that the proposed method works well within the framework of a spatio-temporal data stream.

Title: ***Inference with penalized likelihood for multi-state models with incomplete disease histories in epidemiology***

Speaker: **Ahmadou Alioum**, Inserm, Bordeaux Population Health Research Center, UMR 1219, Univ. Bordeaux, ISPED, F-33000 Bordeaux, France

Schedule: *Monday 11 June, 12:00–12:30*

Abstract: Multi-state models are very useful for modeling the occurrence of several events over time in longitudinal epidemiological studies. The definition of states (often based on plausible clinical or biological conditions) and possible transitions between states depends on the problem under consideration and can lead to very complex models. If, in addition, the observations are made in discrete times so that the exact transition times are unknown, inference for such models becomes very complex. The penalized likelihood approach has been used to estimate transition intensities in multi-state models. One advantage of this approach is that it allows the estimation of smooth transitions intensities in the presence of interval-censored data, without making any parametric assumption on the form of these functions. The main difficulty is the choice of the smoothing parameter which can be done by cross-validation. Another problem is the estimation the variance of the maximum penalized likelihood estimators. The talk will discuss the interest of using penalized likelihood for inference in multi-state models on some examples of models based on longitudinal epidemiological studies.

Title: ***Kernel regression estimation with errors-in-variables for random fields***

Speaker: **Baba Thiam**, LEM, France

Schedule: *Monday 11 June, 12:30–13:00*

Abstract: In this paper, we investigate kernel regression estimation when the data are con- taminated by measurement errors in the context of random fields. We establish sharp rate of weak and strong convergence of the kernel regression estimator under both the ordinary smooth and super-smooth assumptions. Numerical studies were carried out in order to illustrate the performance of the estimator with simulated data.

Title: ***Inference with spatially correlated survival data***

Speaker: **Akim Adekpedjou**, University of Missouri, United States of America

Co-author(s): Sophie Dabo Niang

Schedule: *Monday 11 June, 13:00–13:30*

Abstract: We discuss models for survival data for single and recurrent event that account for spatial correlation and covariates. The models are developed using modern survival analysis techniques coupled with the theory of geo-statistics. Extension to more complex type of data such as error prone and functional covariates are presented. Generalization to multiple units per regions in single and recurrent events cases will be discussed. Some simulation studies results as well as application to illustrate one of the models is presented.

Invited Session IS-03 • Monday 11, 11:30-13:30 • Room Furore

PERMUTATION TESTS FOR FUNCTIONAL AND OBJECT DATA

Organizer(s): Simone Vantini (Politecnico di Milano)

Chair: Simone Vantini (Politecnico di Milano)

Title: ***Local inference for functional data controlling the family-wise error rate on domain subsets***

Speaker: **Alessia Pini**, Department of Statistics, Umeå School of Business and Echonimics, Umeå University, Umeå, Sweden; Department of Statistical Sciences, Catholic University of Sacred Heart, Milano, Italy

Co-author(s): Konrad Abramowicz, Lina Schelin, Sara Sjostedt de Luna, Aymeric Stamm, Simone Vantini

Schedule: *Monday 11 June, 11:30–12:00*

Abstract: In the framework of functional data analysis, inference approaches are two-fold: global inference aiming at testing functional hypotheses over the entire domain and local inference aiming at selecting domain subsets responsible for the rejection of a null hypothesis. In the local setting, one can compute a p-value at every point of the domain, obtaining an unadjusted p-value function, which controls the probability of type I error in a pointwise fashion. Hence, it cannot be used for domain selection purposes as it does not provide any control of the probability of wrongly selecting entire regions. Hence, adjusted p-value functions are needed. In particular, it is often desirable to control the probability of falsely rejecting the null

hypothesis in at least one point of the domain, so-called familywise error rate (FWER). A general family of methods that fit this purpose is discussed. It includes and extends existing methods and our own proposed one. Their inferential properties are characterized in terms of finite-sample or asymptotic control of the FWER and consistency. Finite-sample properties are further compared on a simulation study. Finally, the proposed local inferential techniques are applied to knee kinematic and brain tractography data.

Title: ***Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology***

Speaker: **Alessandra Cabassi**, University of Cambridge, UK

Co-author(s): Davide Pigoli, Piercesare Secchi, Patrick A. Carter, Matteo Fontana, Alessio Farcomeni

Schedule: *Monday 11 June, 12:00–12:30*

Abstract: Pigoli et al (2014) developed a framework to perform inference on the covariance operators of functional random processes, extending distances between positive-definite symmetric matrices to the infinite-dimensional framework. This allowed them to build a permutation test for the equality of the covariance operators of two groups of data. In this work, we generalise the test to the case of multiple groups, using the non-parametric combination methodology of Pesarin and Salmaso (2010) to combine pairwise comparisons into a global test. We consider different combining functions and permutation strategies and explain how it is possible to identify the pairs of groups that have different covariances, when the null hypothesis is rejected. In particular we show that, for some combining functions, step-down adjusting procedures are available to control for multiple testing. The empirical power of the test is then explored via simulations and compared with those of existing alternative approaches in different scenarios. The proposed methodology is applied to data from wheel running activity experiments, that used selective breeding to study the evolution of locomotor behaviour in mice. I will also briefly present an extension of this method to network data and its application to brain imaging, developed during the collaborative StartUpResearch project.

Title: ***Robust testing of generalized spatial regression models by conditional score test***

Speaker: **Livio Finos**, UniPD, Italy

Co-author(s): Jesse Hemerik, Federico Ferraccioli, Laura Sangalli, Jelle Goeman

Schedule: *Monday 11 June, 12:30–13:00*

Abstract: Matthieu and Sangalli (2016, Generalized spatial regression with

differential regularization, Journal of Statistical Computation and Simulation) proposed a generalized additive model for geostatistical and areal data observed over irregularly shaped spatial domains. Despite the model allows to account for spatially varying covariate information, it lacks on an inferential approach to test the effect of these covariates. Due to presence of a penalization term and the failure of maximum likelihood assumptions, the traditional tests are not applicable. Here we accomplish the task by extending the approach of Hemerik, Goeman and Finos (2018, Robust testing in generalized linear models by sign-flipping score contributions, Submitted). This approach is based on random sign-flipping of score contributions. As long as the spatial model is correct on average and under mild assumptions, the method is robust (i.e. asymptotically exact, consistent) against several types of model misspecification, such as overdispersion, heteroscedasticity and, in some cases, ignored nuisance parameters. An application to real data is shown and discussed.

Title: ***Group comparison for network-valued data with an application to the Human brain***

Speaker: **Ilenia Lovato**, Department of Mathematics, University of Pavia, Italy
Co-author(s): Alessia Pini, Aymeric Stamm, Simone Vantini

Schedule: *Monday 11 June, 13:00–13:30*

Abstract: In the context of Object Oriented Data Analysis, we consider network-valued data: the unit of the statistical analysis is a network and one has to deal with samples of networks. Although network analysis has been deeply studied in the literature, statistical methods applicable to samples of networks are almost lacking. Networks are often modelled as objects in non-Euclidean spaces, therefore a generalization of classical Euclidean statistical tools is needed. Two-sample null hypothesis (global or local) testing is one of the first steps towards the construction of a statistical framework for network-valued data. Due to the potential complexity of the data (e.g. non-tractable probabilistic generative models), we propose to approach the problem from the perspective of permutation theory. Exactness and consistency of the resulting tests are shown. The range of possible applications of such tools is very vast, since in a great number of data sets, the statistical unit can be represented as a network. Neuroimaging is without a doubt one of the most important fields where the representation of brains as networks can be groundbreaking. An application of our methodology to study the structure and function of the Human brain will be presented.

Invited Session IS-04 • Monday 11, 11:30-13:30 • Room Positano

MODERN BAYESIAN NON-PARAMETRIC METHODS FOR BIOMEDICAL STUDIES

Organizer(s): Debajyoti Sinha (Florida State University)
Chair: Debajyoti Sinha (Florida State University)

Title: ***A flexible cure rate model for spatially correlated survival data based on generalized extreme value distribution and Gaussian process priors***

Speaker: **Dipak Kumar Dey**, University of Connecticut, United States of America

Schedule: *Monday 11 June, 11:30–12:00*

Abstract: Our present work proposes a new survival model in a Bayesian context to analyze right-censored survival data for populations with a surviving fraction, assuming that the log failure time follows a generalized extreme value distribution. Many applications require a more flexible modeling of covariate information than a simple linear or parametric form for all covariate effects. It is also necessary to include the spatial variation in the model, since it is sometimes unexplained by the covariates considered in the analysis. Therefore, the nonlinear covariate effects and the spatial effects are incorporated into the systematic component of our model. Gaussian processes (GPs) provide a natural framework for modeling potentially nonlinear relationship and have recently become extremely powerful in nonlinear regression. Our proposed model adopts a semiparametric Bayesian approach by imposing a GP prior on the nonlinear structure of continuous covariate. With the consideration of data availability and computational complexity, the conditionally autoregressive distribution is placed on the region-specific frailties to handle spatial correlation. The flexibility and gains of our proposed model are illustrated through analyses of simulated data examples as well as a data set involving a colon cancer clinical trial from the state of Iowa.

Title: ***Bayesian nonparametric inference for the covariate-adjusted ROC curve***

Speaker: **Vanda Inacio de Carvalho**, University of Edinburgh, United Kingdom

Schedule: *Monday 11 June, 12:00–12:30*

Abstract: Accurate diagnosis of disease is of fundamental importance in clinical practice and medical research. Before a medical diagnostic test

is routinely used in practice, its ability to distinguish between diseased and nondiseased states must be rigorously assessed through statistical analysis. The receiver operating characteristic (ROC) curve is the most popular used tool for evaluating the discriminatory ability of continuous-outcome diagnostic tests. Recently, it has been acknowledged that several factors (e.g., subject-specific characteristics, such as age and/or gender) can affect the test's accuracy beyond disease status. In this work, we develop Bayesian nonparametric inference, based on a combination of dependent Dirichlet process mixture models and the Bayesian bootstrap, for the covariate-adjusted ROC curve (Janes and Pepe, 2009, Biometrika), a measure of covariate-adjusted diagnostic accuracy. Applications to simulated and real data are provided.
Joint work with M.X. Rodríguez-Álvarez, BCAM, Spain

Title: ***Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity***
Speaker: **Antonio Linero**, Florida State University, United States of America
Co-author(s): Yun Yang
Schedule: *Monday 11 June, 12:30–13:00*
Abstract: Ensembles of decision trees are a useful tool for obtaining for obtaining flexible estimates of regression functions. Examples of these methods include gradient boosted decision trees, random forests, and Bayesian CART. Two potential shortcomings of tree ensembles are their lack of smoothness and vulnerability to the curse of dimensionality. We show that these issues can be overcome by instead considering sparsity inducing soft decision trees in which the decisions are treated as probabilistic. We implement this in the context of the Bayesian additive regression trees framework, and illustrate its promising performance through testing on benchmark datasets. We provide strong theoretical support for our methodology by showing that the posterior distribution concentrates at the minimax rate (up-to a logarithmic factor) for sparse functions and functions with additive structures in the high-dimensional regime where the dimensionality of the covariate space is allowed to grow near exponentially in the sample size. Our method also adapts to the unknown smoothness and sparsity levels, and can be implemented by making minimal modifications to existing BART algorithms.

Title: ***Bayesian analysis of monotone single index model for mental health data***
Speaker: **Debajyoti Sinha**, Florida State University, United States of America
Schedule: *Monday 11 June, 13:00–13:30*

Abstract: For many mental health studies with unknown non-linear relationship between the response and its multiple predictors, the single index model is a practical dimension reduction tool with good physical interpretation. Major impediments for widespread uses of existing Bayesian analysis for such models include slow mixing of the Markov Chain Monte Carlo (MCMC), inability to deal with bounded and skewed response and a lack of theoretical justification of the convergence of the future prediction. We present a new Bayesian single index model with associated MCMC algorithm that incorporates an efficient Metropolis Hastings (MH) step using a mode-aligned proposal density for the conditional distribution of the index vector. Our method leads to a model with great clinical interpretation and prediction, and implementable Bayesian inference with fast convergence of the MCMC. We also obtain for the first time, the set of sufficient conditions for obtaining optimal rate of convergence of the overall regression function. We illustrate the practical advantages of our method and computational tool via re-analysis of a mental health study.

Invited Session IS-05 • Monday 11, 11:30-13:30 • Room Procida

ADVANCES OF STATISTICAL INFERENCE IN HIGH DIMENSIONAL DATA WITH COMPLEX STRUCTURES

Organizer(s): Wen Zhou (Colorado State University)
Chair: Wen Zhou (Colorado State University)

Title: ***A Unified Nonparametric Procedure on Detecting Spurious Discoveries for Sparse Signals with Diverging Dimensions***
Speaker: **Chao Zheng**, Lancaster University, United Kingdom
Co-author(s): Lvou Zhang, Wen Zhou, Wen-Xin Zhou Zhou
Schedule: *Monday 11 June, 11:30–12:00*
Abstract: Identifying a subset of response-associated covariates from a large number of candidates has become a fundamental tool for scientific discoveries, particularly in biology including the differential analysis in genomics and the genome-wide association study in genetics. However, given the high dimensionality and the sparsity of signals in data, spurious discoveries can easily arise. We introduce a statistical measure on the goodness of spurious fit based on the maximum rank correlations among predictors and responses. The proposed statistic imposes no assumptions on the data types, dependency,

and the underlying models. We derive the asymptotic distribution of such goodness of spurious fit under very mild assumptions on the associations among predictors and responses. We propose a multiplier bootstrap procedure to estimate such a distribution and utilize it as the benchmark to guard against spurious discoveries. It is also applied to the variable selection problems for the high dimensional generalized regressions. We applied our method to genetic studies to demonstrate that the proposed measure provides a statistical verification of the detected biomarkers.

Title: ***Data integration for the simultaneous estimation of normal means***
Speaker: **Sihai Dave Zhao**, University of Illinois at Urbana-Champaign, United States of America
Schedule: *Monday 11 June, 12:00–12:30*
Abstract: The integrative analysis of disparate datasets is an important strategy in data analysis. It is increasingly popular in the field of genomics, which enjoys a wealth of publicly available datasets that can be compared and contrasted, or combined with new data, to extract novel scientific insights. This paper studies a simple but non-trivial example of data integration: leveraging an auxiliary sequence of side information for the simultaneous estimation of a vector of normal means. This task is formulated as a compound decision problem, an oracle integrative decision rule is derived, and a data-driven estimate of this rule, based on minimizing a SURE estimate of the oracle risk, is proposed. The data-driven rule is shown to asymptotically achieve the minimum possible risk among all separable decision rules, and its good performance is demonstrated in numerical properties. The proposed method leads naturally to an integrative high-dimensional classification procedure, which is shown to be capable of outperforming non-integrative methods in problems in genomics.

Title: ***Post-Regularization Confidence Bands for High-Dimensional Non-parametric Models with Local Sparsity***
Speaker: **Mladen Kolar**, University of Chicago, United States of America
Schedule: *Monday 11 June, 12:30–13:00*
Abstract: We propose a novel high dimensional nonparametric model named ATLAS which is a generalization of the sparse additive model. The ATLAS model assumes the high dimensional regression function can be locally approximated by a sparse additive function, while such an approximation may change from the global perspective. We aim to estimate high dimensional function using a novel kernel-sieve hybrid regression estimator that combines the local kernel regression with

the B-spline basis approximation. We show the estimation rate of true function in the supremum norm. We also propose two types of confidence bands for true function. Both procedures proceed in two steps: (1) a novel bias correction method is applied to remove the shrinkage introduced by the model selection penalty and (2) quantiles of the normalized de-biased estimator are approximated by quantiles of the limiting distribution or a Gaussian multiplier bootstrap. We further show that the covering probability of the bootstrap confidence bands converges to the nominal one at a polynomial rate.
Joint work with Junwei Lu and Han Liu.

Title: ***Covariate-adjusted semiparametric latent graphical model***
Speaker: **Wen Zhou**, Colorado State University, United States of America
Co-author(s): Zhao Ren, Hui Zou
Schedule: *Monday 11 June, 13:00–13:30*
Abstract: Motivated from the study on protein folding in structure biology, we propose a covariate-adjusted latent semiparametric graphical model for modeling the
conditional independency among a set of amino acids adjusting for possible
external effects. Different from other studies in system biology, data from protein folding analysis usually are of count type with a small number of unique values and subjects to external effects such as contact group volumes and energy levels. Using the Gaussian copula distribution, the proposed model provides a novel approach to estimate the conditional independence relationship among latent variables based on the observed count data after the confounding effects are taken into account. Our work has two main theoretical contributions: (1) a smoothed rank correlation maximization procedure is proposed to adjust the confounding effect and a uniform error bound is established; (2) using the rank-based approach, the latent conditional independence relationship is estimated and the convergence rate is established for graph recovery. The proposed model and method are numerically assessed through extensive simulation studies. The model is applied to a real study for detecting co-evolution signals from multiple sequence alignment data, and reveals interesting biology insights.

NONPARAMETRIC TEST AND MODELLING I

Chair: Geoffrey Decrouez (Higher School of Economics)

Title: ***Standard Testing Procedures for White Noise and Heteroskedasticity***

Speaker: **Violetta Dalla**, National and Kapodistrian University of Athens

Co-author(s): Liudas Giraitis, Peter C. B. Phillips

Schedule: *Monday 11 June, 15:00–15:20*

Abstract: Commonly used tests to assess evidence for the absence of serial correlation between time series in applied work 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 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. 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.

Title: ***Stratified permutation tests with applications to University of Padova students' careers***

Speaker: **Eleonora Carrozzo**, University of Padova, Italy

Co-author(s): Rosa Arboretti, Luigi Salmaso

Schedule: *Monday 11 June, 15:20–15:40*

Abstract: The problem of detecting differences between two stratified groups is quite common in the field of multi-clinical trials. In relatively recent years a multivariate extension for categorical variables of the well known van Elteren test has been proposed in literature showing good operating characteristics. But for most of the sample sizes adopted in practice (?10) this test is biased. In the present work, we present a nonparametric procedure, based on permutation test and Nonparametric Combination (NPC) methodology allowing us to deal with relatively small samples regardless the type of underling distribution. Furthermore, we combine this procedure with a method to test certain order relations on the distribution of the response

variables. We show the application of the proposed procedure to data of Engineering students' careers at the University of Padova.

Title: ***Non-Parametric Recursive Estimation of the Copula***
Speaker: **Geoffrey Decrouez**, Higher School of Economics, Russian Federation
Co-author(s): Felix Camirand
Schedule: *Monday 11 June, 15:40–16:00*
Abstract: Traditional non-parametric estimators of a copula, such as the ones based on empirical quantiles or on pseudo observations, must be entirely recalculated as new observations are collected. We introduce two new non-parametric recursive estimators of the copula, and of the copula density. Recursive estimators are those that can be updated in $O(1)$, so that the total cost after n observations are received is $O(n)$. All estimators require a recursive estimation of the quantile, which is achieved using a stochastic approximation algorithm. We derive the asymptotic normal distribution of the recursive estimators, and we illustrate their performance in a numerical study.

Contributed Session CS-06 • Monday 11, 15:00-16:00 • Room Furore

DENSITY ESTIMATION

Chair: Maarten Jansen (Université libre de Bruxelles)

Title: ***Adaptive unfolding spheres' size distribution from linear sections.***
Speaker: **Zbigniew Szkutnik**, AGH University of Science and Technology, Krakow, Poland
Co-author(s): Bogdan Ćmiel, Jakub Wojdyła
Schedule: *Monday 11 June, 15:00–15:20*
Abstract: The stereological Spektor-Lord-Willis problem of unfolding the distribution of spheres radii from linear sections through an opaque medium containing the spheres is formulated as a Poisson inverse problem and an L2-rate-minimax solution is constructed over some restricted Sobolev classes. The solution is a specialized kernel-type estimator with boundary correction. Non-parametric, asymptotic confidence bands for the unfolded function are constructed. Automatic bandwidth selection procedures based on empirical risk minimization are proposed. It is shown that a version of the Goldenshluger-Lepski procedure ensures adaptivity of the estimators to the unknown smoothness in the scale of Sobolev classes. The performance of the procedures is demonstrated in a Monte Carlo experiment. It should be

stressed that, in contrast to all the previous constructions of confidence bands in inverse problems, our bands are constructed in the Poisson inverse problem setup with random number of data points, which is more realistic in stereology than inverse density estimation setup with non-random sample size.

Title: ***Mellin-Meijer kernel density estimation on R+***
Speaker: **Gery Geenens**, UNSW Sydney, Australia
Schedule: *Monday 11 June, 15:20–15:40*
Abstract: Nonparametric kernel density estimation is a very natural procedure which simply makes use of the smoothing power of the convolution operation. Yet, it performs poorly when the density of a positive variable is to be estimated (boundary issues, spurious bumps in the tail). So, various extensions of the kernel estimator allegedly suitable for $\mathfrak{R}^+$-supported densities, such as those using asymmetric kernels, abound in the literature. Those, however, are not based on any valid smoothing operation analogous to the convolution, which typically leads to inconsistencies. By contrast, in this paper a kernel estimator for $\mathfrak{R}^+$-supported densities is defined by making use of the Mellin convolution, the natural analogue of the usual convolution on $\mathfrak{R}^+$. From there, a very transparent theory flows and leads to a class of asymmetric kernels strongly related to Meijer G-functions. Numerous pleasant properties of this 'Mellin-Meijer-kernel density estimator' are presented. Its pointwise- and L_2-consistency (with optimal rate of convergence) are established for a large class of densities, including densities unbounded at 0 and showing power-law decay in their right tail. Its practical behaviour is investigated further through simulations and some real data analyses.

Title: ***Density estimation using multiscale local polynomial transforms***
Speaker: **Maarten Jansen**, Université libre de Bruxelles, Belgium
Schedule: *Monday 11 June, 15:40–16:00*
Abstract: The estimation of a density function with an unknown number of singularities or discontinuities is a typical example of a multiscale problem, with data observed at nonequispaced locations. The data are analysed through a multiscale local polynomial transform (MLPT), which can be seen as a slightly overcomplete, non-dyadic alternative for a wavelet transform, equiped with the benefits from a local polynomial smoothing procedure. In particular, the multiscale transform adopts a sequence of kernel bandwidths in the local polynomial smoothing as resolution level dependent, user controlled scales. The MLPT analysis leads to a reformulation of the problem as a variable selection in a sparse, high-dimensional

regression model with exponentially distributed responses. The variable selection is realised by the optimisation of the l1-regularised maximum likelihood, where the regularisation parameter acts as a threshold. Fine-tuning of the threshold requires the optimisation of an information criterion such as AIC. Both theoretical results and simulation studies are presented.

Contributed Session CS-07 • Monday 11, 15:00-16:00 • Room Positano

INFERENCE IN HIGH DIMENSIONS

Chair: Eugen Pircalabelu (KU Leuven)

Title: ***Improved bounds for Square-Root Lasso and Square-Root Slope***

Speaker: **Alexis Derumigny**, CREST-ENSAE, France

Schedule: *Monday 11 June, 15:00–15:20*

Abstract: Extending some results to the setting of sparse high-dimensional linear regression with unknown variance, we show that two estimators, the Square-Root Lasso and the Square-Root Slope can achieve the optimal minimax prediction rate, which is $(s/n)\log(p/s)$, up to some constant, under some mild conditions on the design matrix. Here, n is the sample size, p is the dimension and s is the sparsity parameter. We also prove optimality for the estimation error in the l_q-norm, with $q \in [1,2]$ for the Square-Root Lasso, and in the l_2 and sorted l_1 norms for the Square-Root Slope. Both estimators are adaptive to the unknown variance of the noise. The Square-Root Slope is also adaptive to the sparsity s of the true parameter. Next, we prove that any estimator depending on s which attains the minimax rate admits an adaptive to s version still attaining the same rate. We apply this result to the Square-root Lasso. Moreover, for both estimators, we obtain valid rates for a wide range of confidence levels, and improved concentration properties where the case of known variance is treated. Our results are non-asymptotic.

Title: ***Correction for optimisation bias in structured sparse high-dimensional variable selection***

Speaker: **Bastien Marquis**, Université libre de Bruxelles, Belgium

Co-author(s): Maarten Jansen

Schedule: *Monday 11 June, 15:20–15:40*

Abstract: In sparse high-dimensional data, the selection of a model can lead to an overestimation of the number of nonzero variables. Indeed, the use of an l_1 norm constraint while minimising the sum of squared

residuals tempers the effects of false positives, thus they are more likely to be included in the model. On the other hand, an l_0 regularisation is a non-convex problem and finding its solution is a combinatorial challenge which becomes unfeasible for more than 50 variables. To overcome this situation, one can come up with a selection via an l_1 penalisation but estimate its coefficients without shrinkage. This leads to an additional bias in the optimisation of an information criterion over the model size. Used as a stopping rule, this IC must be modified to take into account the deviation of the estimation with and without shrinkage. By looking into the difference between the Prediction Error and the expected Mallows's Cp, previous work analysed a correction for the optimisation bias and an expression can be found for a signal-plus-noise model given some assumptions. A focus on structured models, in particular grouped variables, shows similar results, though the bias is noticeably reduced.

Title: ***High-dimensional multiresolution graphs for functional brain connectivity***

Speaker: **Eugen Pircalabelu**, KU Leuven, Belgium

Co-author(s): Gerda Claeskens, Lourens J. Waldorp

Schedule: *Monday 11 June, 15:40–16:00*

Abstract: A new method is proposed to simultaneously estimate graphical models from data obtained at K different coarseness scales. Starting from a predefined scale $k^* \leq K$, the method offers the possibility to zoom in or out over scales on particular edges. The estimated graphs over the different scales have similar structures, although their sparsity level depends on the scale at which estimation takes place. The graphs are jointly estimated at all coarseness scales and the method makes it possible to evaluate the evolution of the graphs from the coarsest to the finest scale or vice-versa. The method is motivated by fMRI datasets that do not all contain measurements on the same set of brain regions. For certain datasets, some of the regions have been split in smaller subregions and the purpose is to estimate sparse graphical models. We accomplish this by pooling information from all subjects in order to estimate a common undirected and directed graph at each coarseness scale, accounting for time dependencies and multiple coarseness scales and by jointly estimating the graphs at all coarseness scales. Empirical and theoretical evaluations illustrate the usefulness of the method and show the method's performance in practice.

COUNT DATA

Chair: Mehdi Razzaghi (Bloomsburg University)

Title: ***Model evaluation of the Fed monetary rules***
Speaker: **Igor Kheifets**, ITAM, Mexico
Co-author(s): Carlos Velasco
Schedule: *Monday 11 June, 15:00–15:20*
Abstract: In this paper we evaluate different modeling alternatives to describe the decisions of central banks in the control of monetary policy instruments. In particular we concentrate on single and multi-step models to describe the discrete nature of monetary policy rules. We use different goodness-of-fit techniques adapted to the discrete nature of the data and of the decision mechanism. The distinctive feature of our approach is that our test statistics are based on original discrete data that avoids any randomization or use of jittering. We show that the Federal Reserve reaction function is best described by multi-step models that allow for economic variables and previous decisions to condition in different and completely unrestricted ways the decisions of making a movement in the target interest rate, and the sign and magnitude of the decided changes.

Title: ***Semiparametric weighting estimations of a zero-inflated Poisson model with missing covariates***
Speaker: **Martin Tshishimbi Wa Lukusa**, Academia Sinica, Taiwan
Co-author(s): Frederick Phoa Kin Hing
Schedule: *Monday 11 June, 15:20–15:40*
Abstract: We scrutinize the problem of missing covariates under a zero-inflated Poisson regression model. Often in the presence of excess zeroes, the traditional count regressions models may fail to accurately estimate the probability of zero state and the probability of the nonzero state. Moreover, we assume that some covariates which model the probability of zero state and the probability of the nonzero state are missing at random. Under the missing at random, the complete-case estimator is biased and inefficient. Although the inverse probability weighting estimator is unbiased, but still remains less efficient. Parametric augmentation inverse probability weighting estimators are the most used approaches to improve the inefficiency issue. Under the zero-inflated Poisson model framework, we develop four types of robust weighting estimators including two full-based types and two

hybrid types where the selection probabilities and the augmentation parts are estimated by the generalized additive models (GAMs) and the kernel-assisted method. The benefits of using a GAM selection probability or a GAM augmentation term in estimating the weithing estimators of a zero-inflated Poisson model are underlined. Moreover, a Monte Carlo experiment and a specific sample set are used to investigate the merit of the proposed methods.

Title: ***Random Effect Models for Count Data with Underdispersion***
Speaker: **Mehdi Razzaghi**, Bloomsburg University, United States of America
Schedule: *Monday 11 June, 15:40–16:00*
Abstract: Several distributions including the Poisson-gamma, generalized Poisson, and Conway-Maxwell-Poisson have been introduced for modeling count data with overdispersion. Molenburgh et al (2007) extend the Poisson-gamma regression for repeated and clustered overdispersed data. Their approach is based on defining a generalized linear model which accommodates overdispersion and clustering through two separate sets of random effects using Poisson-gamma and normal distributions. Here, we consider the extension of random effect models for underdispersed data. Our approach will be similar to that of Molenburgh et al (2007), but we use the hyper-Poisson distribution. The advantage of this approach is that both overdispersion and underdispersion can be accounted for.

Special Invited Session SIS-02 • Monday 11, 15:00-16:00 • Auditorium Tafuri

NORM ESTIMATION IN SPARSE LINEAR REGRESSION WITH VARIOUS ASSUMPTIONS ON NOISE AND DESIGN MATRIX.

by Alexandre Tsybakov[(1)] **& Olivier Collier**[(2)] **& Laëtitia Comminges**[(1)]
[(1)] University of Paris VI, France; [(2)] Université Paris Nanterre, France

Speaker: Olivier Collier
Chair: Patrice Bertail (Université Paris Nanterre)

Abstract: In the subGaussian sequence model, we prove that minimax noise level estimation cannot be performed at the same rate as in the Gaussian case, and we show that this implies a similarly deprecated rate for norm estimation. We also investigate other phenomenons with different noise distribution and design matrix assumptions.

Invited Session IS-06 • Monday 11, 16:30-18:30 • Room Vietri

RECENT ADVANCES IN COMPLEX DATA ANALYSIS

Organizer(s): Juan Romo (Universidad Carlos III de Madrid)
Chair: Juan Romo (Universidad Carlos III de Madrid)

Title: ***Depth analysis for sparse functional data***
Speaker: **Carlo Sguera**, Universidad Carlos III de Madrid, Spain
Co-author(s): Sara López-Pintado
Schedule: *Monday 11 June, 16:30–17:00*
Abstract: Data depth is a well-known and useful notion in functional data analysis. It provides a center-outward ranking for a sample of curves. This ordering allows the definition of descriptive statistics such as medians, trimmed means and central regions for functional data. Moreover, data depth is often used as a building block for developing outlier detection techniques and for robustifying standard statistical methods.

Functional depths have been originally proposed for sample of curves that are measured on a common and dense grid. In practice, this is not always the case, since curves are often observed at subject-dependent and/or sparse grids. The main approach in the literature for dealing with this situation is based on estimating the individual trajectories on an artificially created common and dense grid of points and using these estimated curves as observed data in the notion of depth. Up to the date, the proposals based on this approach ignored the inherent uncertainty associated to the preliminary curve estimation step.

We propose an alternative approach for functional depths that explicitly address sparsity and we design a general procedure that allows to take uncertainty into account.

Title: ***High-dimensional sign tests for the direction of a skewed single-spiked distribution***
Speaker: **Davy Paindaveine**, Université libre de Bruxelles, Belgium
Co-author(s): Thomas Verdebout
Schedule: *Monday 11 June, 17:00–17:30*
Abstract: We consider the problem of testing the null hypothesis that the direction θ of a skewed single-spiked high-dimensional distribution coincides with a given direction θ_0. For robustness purposes, we restrict to spatial sign tests, that is, to tests that involve the observations only through their projections onto the unit sphere. This reduces the problem to a classical problem in directional statistics,

namely to the spherical location testing problem, for which the Watson test is the standard procedure. We study the asymptotic null and non-null behaviours of this test, in a general asymptotic framework where the dimension converges to infinity in an arbitrary way as a function of the sample size n. We also allow the strength of the signal to behave in a completely free way with n, which provides a complete spectrum of problems ranging from arbitrarily challenging to arbitrarily easy problems. Our results identify several asymptotic regimes leading to different limiting asymptotic experiments. Asymptotically optimal tests are obtained in each regime. Monte Carlo studies support our theoretical results.

Title: ***Statistical Data Depths in Robust Parametric Estimation***
Speaker: **Claudio Agostinelli**, University of Trento, Italy
Schedule: *Monday 11 June, 17:30–18:00*
Abstract: Statistical Data Depths are an important tool in the analysis of data. Their mean use in the non parametric setting. We will discuss different ways of using Depth functions to the aim of obtaining robust parametric estimators. One general approach is to construct Weighted Likelihood Estimating Equations where the weight, attached to each score contribution, is evaluated by comparing the statistical data depth at the model with that of the sample in a given point. Observations are considered regular when the ratio of these two depths is close to one, whereas, when the ratio is large the corresponding score contribution may be downweigthed. Details and examples are provided for the robust estimation of the parameters in the multivariate normal model. Because of the form of the weights, we expect that, there will be no downweighting under the true model leading to highly efficient estimators. Robustness is illustrated using real data sets.

Title: ***Visualization and Assessment of Spatio-temporal Covariance Properties***
Speaker: **Ying Sun**, KAUST, Saudi Arabia
Co-author(s): Huang Huang
Schedule: *Monday 11 June, 18:00–18:30*
Abstract: Spatio-temporal covariances are important for describing the spatio-temporal variability of underlying random fields in geostatistical data. For second-order stationary random fields, there exist subclasses of covariance functions that assume a simpler spatio-temporal dependence structure with separability and full symmetry. However, it is challenging to visualize and assess separability and full symmetry from spatio-temporal observations. In this work, we propose a functional data analysis approach that constructs test

functions using the cross-covariances from time series observed at each pair of spatial locations. These test functions of temporal lags summarize the properties of separability or symmetry for the given spatial pairs. We use functional boxplots to visualize the functional median and the variability of the test functions, where the extent of departure from zero at all temporal lags indicates the degree of non-separability or asymmetry. We also develop a rank-based nonparametric testing procedure for assessing the significance of the non-separability or asymmetry. The performances of the proposed methods are examined by simulations with various commonly used spatio-temporal covariance models. To illustrate our methods in practical applications, we apply it to real datasets, including weather station data and climate model outputs.

Invited Session IS-07 • Monday 11, 16:30-18:30 • Room Ravello

TESTING IN NON AND SEMIPARAMETRIC MODELS

Organizer(s): Pascal Lavergne (Toulouse School of Economics)
Chair: Pascal Lavergne (Toulouse School of Economics)

Title: ***Nonparametric Significance Testing in Measurement Error Models***
Speaker: **Luke Taylor**, Aarhus University, Denmark
Schedule: *Monday 11 June, 16:30–17:00*
Abstract: We develop the first nonparametric significance test for regression models with measurement error in the regressors. Despite our test using deconvolution estimators - with notoriously slow convergence rates - we show it is able to detect local alternatives at the $\sqrt{n}$ rate. We also discuss two empirical applications. The first considers the effect of cognitive ability on a range of socio-economic variables. The second uses time series data - and a novel approach to estimate the measurement error without repeated measurements - to investigate whether future inflation expectations are able to stimulate current consumption.

Title: ***Specification Tests for Semiparametric Conditional Moments with Generated Covariates***
Speaker: **Elia Lapenta**, Toulouse School of Economics, France
Schedule: *Monday 11 June, 17:00–17:30*
Abstract: This paper develops a specification test for a semiparametric conditional-moment restriction involving an unknown finite-dimensional parameter and some covariates which are not observed but are identified as

either the nonparametric conditional expectations of some observed variables, or as the residuals from auxiliary nonparametric regressions.
The test can be used in order to check the specification of semiparametric models with control functions, semiparametric game-theoretical models with incomplete information, semiparametric double-hurdle models, semiparametric sample-selection models with a nonparametric selection equation.
The test we propose is an ICM test, although our proves also hold for a Kolgomorov-Smirnov type test. By using tools from U-Process Theory and Empirical Process Theory we derive its asymptotic distribution. In order to get the critical values, we propose a wild-bootstrap procedure and prove the consistency of the bootstrap test.

Title: ***A Bootstrap Test for Bayesian-Nash Equilibria in Nonparametric Discrete Games***
Speaker: **Pascal Lavergne**, Toulouse School of Economics, France
Co-author(s): Elia Lapenta
Schedule: *Monday 11 June, 17:30–18:00*
Abstract: We develop a nonparametric test for the hypothesis that the conditional expectation of Y given X is equal to the conditional expectation of Y given W(X),where both Y and X are observed, while W(X) is a vector that contains unknown conditional expectations. The test can be used to check whether the distribution of the observed data is generated by a Bayesian-Nash equilibrium of a discrete game with private information and possibly correlated types of agents. The procedure is an Integrated Conditional Moment Test and uses kernel estimators for the unknown W(X). Using empirical process theory, we study the asymptotic properties of our test statistic. Since its asymptotic distribution depends on unknown quantities, bootstrap is used to obtain asymptotically valid critical values. Proofs account for a data-dependent bandwidth.

Title: ***Distributional Regression in Survival Analysis***
Speaker: **Andres García-Suaza**, Universidad del Rosario, Colombia
Schedule: *Monday 11 June, 18:00–18:30*
Abstract: This article proposes inference procedures on distributional regression models in the context of survival analysis under random censoring. The proposal generalizes classical survival models to a situation where slope coefficients depend on survival time. We formally justify asymptotic inference on the varying coefficients under fairly weak regularity conditions, similar to those needed when data is not censored. Finite sample properties of the proposed methods are stud-

ied by means of Monte Carlo experiments. Finally, we apply the proposed methods to study the effect of unemployment benefits on unemployment duration.

Invited Session IS-08 • Monday 11, 16:30-18:30 • Room Furore

NONPARAMETRIC METHODS FOR CENSORED OR TRUNCATED DATA

Organizer(s): Somnath Datta (University of Florida)
Chair: Somnath Datta (University of Florida)

Title: ***A non-parametric test for evaluating censored biomarker data***
Speaker: **María Carmen Pardo**, Complutense University of Madriid, Spain
Co-author(s): Alba María Franco-Pereira, Christos Nakas
Schedule: *Monday 11 June, 16:30–17:00*
Abstract: In practice, biomarkers are often imprecisely measured due to the limitation of assay sensitivity. The values below the limit of detection are not detectable. Ignorance of such data characteristic may lead to inaccurate estimation of marker's potential discriminatory power. The assessment of a diagnostic marker when marker measurements are censored is addressed. We adapt a non-parametric likelihood ratio two-sample test for right censored data and compare with different parametric approaches based on the ROC curve and the ROC/AUC thorough a Monte Carlo simulation study. Finally, we illustrate our approach with a real data set.

Title: ***Estimation of the cumulative incidence function for paired data***
Speaker: **Thomas Scheike**, University of Copenhagen, Denmark
Schedule: *Monday 11 June, 17:00–17:30*
Abstract: In this talk I will discuss how to estimate different useful quantities such as the cumulative incidence function, in the context of paired competing risks data. First of all the marginal cumulative incidence can be estimated using the usual estimator and robust standard errors. In addition one can use random effects models to estimate also a dependence parameter. In addition for paired data, for example based on twins, one can also consider the cumulative incidence conditional on information about the other twin and compare to the marginal cumulative incidence to learn about possible dependence in the data.

Title: ***Recent advances in nonparametric estimation from doubly truncated lifetimes***
Speaker: **Jacobo de Uña-Alvarez**, University of Vigo, Spain

Schedule: *Monday 11 June, 17:30–18:00*

Abstract: Doubly truncated lifetimes often appear in Survival Analysis and Epidemiology, among other fields. Under double truncation, the observed lifetimes belong to a random interval which may vary from individual to individual. The naive analysis of such lifetime data can be systematically biased due to the double truncation issue, so suitable corrections are needed. The seminal paper of Efron and Petrosian (J Amer Statist Assoc, 1999) gave rise to a number of developments in this area of reasearch. However, several technical and practical issues remain still unsolved, and the sampling scheme itself and the corresponding potential biases are often overlooked. In this work we review some recent advances in nonparametric estimation from doubly truncated data, including: estimation of the lifetime cumulative distribution, correlation analysis, regression and multi-state models. Simulations and real data illustrations will be provided, as well as discussion of the asymptotic theory needed for inference purposes.

Title: ***Adjustments of Mann-Whitney U-Statistics for Comparing Sojourn Time Distributions in Observational Studies When Transition Time are Right Censored***

Speaker: **Somnath Datta**, University of Florida, United States of America

Co-author(s): Yichen Chen

Schedule: *Monday 11 June, 18:00–18:30*

Abstract: We propose adjustments to the classical two-sample Mann-Whitney U-Statistics when distributions of sojourn time are compared in an observational study on a multistate system. Our adjustments are geared towards removing potential bias come from right censoring of transition times (state entry or exist times) and presence of confounding covariates. The Mann-Whitney U-statistics are re-weighted and normalized based on a combination of inverse probability of censoring weights and propensity score based weights. Two types of the adjustments are proposed corresponding to two different selection methods. The dependence of censoring time on state entry time, group membership and/or observed covariates is considered. Simulation results are used to illustrate the performance of the adjusted Mann-Whitney U-statistics under different censoring mechanisms and imputation approaches. A real data set on recovery process following bone marrow transplant for acute leukemia was applied to exemplify its usefulness.

Invited Session IS-09 • Monday 11, 16:30-18:30 • Room Positano

NON PARAMETRIC AND STATISTICAL LEARNING CHALLENGES FOR BIG DATA

Organizer(s): Patrice Bertail (Université Paris Nanterre)
Chair: Patrice Bertail (Université Paris Nanterre)

Title: ***On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability***
Speaker: **Stephan Clémençon**, Telecom ParisTech, France
Co-author(s): Guillaume Papa, Aurélien Bellet
Schedule: *Monday 11 June, 16:30–17:00*
Abstract: The problem of predicting connections between a set of data points finds many applications, in systems biology and social network analysis among others. This paper focuses on the *graph reconstruction* problem, where the prediction rule is obtained by minimizing the average error over all n(n-1)/2 possible pairs of the n nodes of a training graph. Our first contribution is to derive learning rates of order $O(\log n/n)$ for this problem, significantly improving upon the slow rates of order $O(1/\sqrt{n})$ established in the seminal work of Biau & Bleakley (2006). Strikingly, these fast rates are universal, in contrast to similar results known for other statistical learning problems (e.g., classification, density level set estimation, ranking, clustering) which require strong assumptions on the distribution of the data. Motivated by applications to large graphs, our second contribution deals with the computational complexity of graph reconstruction. Specifically, we investigate to which extent the learning rates can be preserved when replacing the empirical reconstruction risk by a computationally cheaper Monte-Carlo version, obtained by sampling with replacement $B \ll n^2$ pairs of nodes. Finally, we illustrate our theoretical results by numerical experiments on synthetic and real graphs.

Title: ***Predicting plant endemicity based on herbarium data: application to French data***
Speaker: **Jessica Tressou**, UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France
Schedule: *Monday 11 June, 17:00–17:30*
Abstract: Evaluating formal threat criteria for every organism on earth is a tremendously resource-consuming task which will need many more years to accomplish at the actual rate. We propose here a method

allowing for a faster and reproducible threat prediction for the 360,000+ known species of plants. Threat probabilities are estimated for each known plant species through the analysis of the data from the complete digitization of the largest herbarium in the world using machine learning algorithms, allowing for a major breakthrough in biodiversity conservation assessments worldwide. First, the full scientific names from Paris herbarium database were matched against all the names from the international plant list using a text mining open source search engine called Terrier. A series of statistics related to the accepted names of each plant were computed and served as predictors in a statistical learning model with a binary output. The training data was build based on the International Union for Conservation of Nature global Redlisting plants assessments. For each accepted name, the probability to be LC was estimated with a confidence interval and a global misclassification rate of 20%. Results are presented on the world map and according to different plant traits.

Title: ***Semantic embedding and keywords clustering in Google Adwords***
Speaker: **Pietro Fodra**, Cdiscount, France
Schedule: *Monday 11 June, 17:30–18:00*
Abstract: When a Google user performs a query, the engine answers with a multi-results page. The body of the page is populated by results that Google identifies as pertinent, but it can also show commercial Ads pushed by e-commerce actors. For each query, Google creates an auction among vendors: their ad presence and position in the results page depend on the content relevance and bid level. So, how to determine the vendor optimal bid? This question has intuitive answers when a query has a lot of historical signal, but becomes trickier for less popular keywords. An answer is provided by exploring keywords semantics via an NLP algorithm (Word2Vec) trained on e-commerce and generic corpora, which embeds the query space into a Euclidean space equipped with a (semantic) cosine distance, that can be compared to a behavioral one induced by the Jacquard index. In order to mutualize the historical signal on the average return per click estimator, multiple strategies are chosen: the first is clustering the keywords in their Euclidean space, the second is using a weighted KNN. An actual A/B test on Adwords campaigns shows the potential of this approach on the keywords from the long tail.

Title: ***Nonnegative Matrix Factorization: a (Semi-)Parametric Statistical View***
Speaker: **Melanie Zetlaoui**, Paris Nanterre University, France

Co-author(s): Patrice Bertail, Stephan Clémençon
Schedule: *Monday 11 June, 18:00–18:30*
Abstract: It is the purpose of this paper to investigate the *Nonnegative Matrix Factorization* (NMF) task from a statistical semiparametric perspective. Stated in geometrical terms, the goal of NMF consists in finding a convex cone $\mathscr{C}$ of dimension $K \geq 1$ "representing accurately" a cloud of n multivariate data $V = (v_1, \ldots, v_n)$ in the positive orthant $\mathbb{R}_+^F$, with $K \leq F$. Whereas the majority of the literature dedicated to NMF focused on algorithmic issues related to the computation of representations maximizing some goodness-of-fit criterion, statistical grounds for such M-estimation techniques have not been exhibited yet. We formulate the NMF, in a semi-parametric context, as a latent variable model. In this context, we compute the different tangent spaces as well as the efficient score function and propose a Z-estimator with estimated nuisance parameters based on the efficient score. Under appropriate assumptions, this Z-estimator yields asymptotically normal estimates of $\mathscr{C}$'s rays.In a parametric context, model selection issues related to the dimension of the underlying cone $\mathscr{C}$ are also considered through the AIC and BIC approaches.

Invited Session IS-10 • Monday 11, 16:30-18:30 • Room Maiori

NEW DEVELOPMENTS IN MULTIVARIATE INFERENCE

Organizer(s): Arne Bathke (University of Salzburg)
Chair: Arne Bathke (University of Salzburg)

Title: ***A Fast and Robust Way to Estimate Overlap of Niches, and Draw Inference***
Speaker: **Judith Helen Parkinson**, Paris-Lodron University Salzburg, Austria
Co-author(s): Raoul Kutil, Jonas Kuppler, Robert R. Junker, Wolfgang Trutschnig, Arne C. Bathke
Schedule: *Monday 11 June, 16:30–17:00*
Abstract: Recent methodological progress has reignited interest in the problem of quantifying niche overlap. In ecology, not only the quantification of species niches is of interest but also the quantification of the multivariate space of a community, among others. Another application area would be given in economics where one may be interested in the market niche of a company or a product. A long time there were no adequate methods for the quantification, until

a generalized approach was proposed by Blonder in 2014, other methods followed.
We provide an interpretable quantification of niche overlap without imposing a particular parametric model or distribution family. That is, the approach presented here is fully non-parametric. The basis for our approach is given by the recently published non-parametric solution approach which consists of dynamic range boxes. Our method uses rank statistics for the estimators of intervals wherefore it provides a quicker calculable and easier interpretable approach. Further, we provide confidence intervals for the true overlap which have not been yet provided in any of the papers which estimate niche overlap.

Title: ***Inference in a general multivariate analysis of covariance model using a model-free bootstrap technique***

Speaker: **Georg Zimmermann**, Department of Mathematics, Paris Lodron University, Salzburg, Austria.; Department of Neurology, Christian Doppler Clinic and Centre for Cognitive Neuroscience, Paracelsus Medical University, Salzburg, Austria.; Spinal Cord Injury and Tissue Regeneration Centre, Paracelsus Medical University, Salzburg, Austria.

Co-author(s): Markus Pauly, Arne C. Bathke

Schedule: *Monday 11 June, 17:00–17:30*

Abstract: In medical research, it is often necessary to account for one or several covariates when comparing the means of two or more groups. This can be done by assuming a uni- or multivariate analysis of covariance (ANCOVA) model for the data. However, "classical" parametric tests are based on quite restrictive assumptions (homoskedasticity and normality of the errors), which cannot be checked meaningfully in small sample size settings. Therefore, a general multivariate ANCOVA model is being considered, allowing for heteroskedasticity as well as non-normality. Heteroskedasticity-consistent covariance matrix estimation techniques are employed, and a Wald-type as well as an ANOVA-type statistic for testing hypotheses about the adjusted mean vectors are proposed. Finite-sample properties of the corresponding asymptotic tests are examined in a simulation study. Moreover, a model-free bootstrap technique is also applied, in order to improve the type I error control in small samples. To conclude, potential real-life applications in research on rare diseases and preclinical experiments are discussed.

Title: ***On a new efficiency for the comparison of confidence regions***

Speaker: **Daniel Gaigall**, Leibniz Universität Hannover, Germany

Co-author(s): Ludwig Baringhaus

Schedule: *Monday 11 June, 17:30–18:00*

Abstract: We introduce a new asymptotic relative efficiency for the comparison of different confidence regions for an unknown multivariate parameter on the basis of the expected volume. For a class of confidence regions, obtained by families of tests via the correspondence principle, the behavior of the expected volume for increasing sample size is described in terms of the (efficient) Fisher information matrix. The class includes simple confidence ellipsoids and confidence regions based on families of likelihood ratio and Wald type tests. The results apply, e.g., for multinomial distributions. We consider confidence regions for the expectation vector of a multivariate normal distribution and for the attributable risk in a contingency table as examples.

Title: ***Rank-based and Resampling-based Inference Methods for Data with Multiple Endpoints***

Speaker: **Arne Bathke**, University of Salzburg, Austria

Schedule: *Monday 11 June, 18:00–18:30*

Abstract: When there are several endpoints and different predictors, researchers typically want to find out which predictors are relevant, and for which endpoints. We present two rather general approaches trying to accomplish these goals, accommodating binary, ordinal, and metric endpoints, and different nominal factors. One of them uses rank-based statistics and an F-approximation of the sampling distribution, the other uses asymptotically valid resampling techniques. We also try to address the question of how well the proposed methods actually accomplish their goals.

Invited Session IS-11 • Monday 11, 16:30-18:30 • Room Amalfi

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL AND FUNCTIONAL TIME SERIES

Organizer(s): Holger Dette (Ruhr-Universität Bochum)

Chair: Carsten Jentsch (TU Dortmund)

Title: ***A Nonparametric Eigenvalue-Regularized Integrated Covariance Matrix Estimator for Asset Return Data***

Speaker: **Clifford Lam**, London School of Economics and Political Science, United Kingdom

Co-author(s): Phoenix Feng

Schedule: *Monday 11 June, 16:30–17:00*

Abstract: In high-frequency data analysis, the extreme eigenvalues of a

realized covariance matrix are biased when its dimension p is large relative to the sample size n. Furthermore, with non-synchronous trading and contamination of microstructure noise, we propose a nonparametrically eigenvalue-regularized integrated covariance matrix estimator (NERIVE) which does not assume specific structures for the underlying integrated covariance matrix. We show that NERIVE is positive definite in probability, with extreme eigenvalues shrunk nonlinearly under the high dimensional framework $p/n \to c > 0$. We also prove that in portfolio allocation, the minimum variance optimal weight vector constructed using NERIVE has maximum exposure and actual risk upper bounds of order $p^{-1/2}$. Incidentally, the same maximum exposure bound is also satisfied by the theoretical minimum variance portfolio weights. All these results hold true also under a jump-diffusion model for the log-price processes with jumps removed. They are further extended to accommodate the existence of pervasive factors such as a market factor under the setting $p^{3/2}/n \to c > 0$. The practical performance of NERIVE is illustrated by comparing to the usual two-scale realized covariance matrix as well as some other nonparametric alternatives using different simulation settings and a real data set.

Title: ***Estimating the eigenvalue spectrum of (auto)covariance matrices of high-dimensional time series***

Speaker: **Alexander Aue**, University of California, Davis, United States of America

Co-author(s): Jamshid Namdari, Debashis Paul

Schedule: *Monday 11 June, 17:00–17:30*

Abstract: This talk is concerned with developing estimation for a class of high-dimensional time series by means of estimating eigenvalue distributions of their coefficient matrices. To do so, a Marcenko-Pastur-type limit distribution for a weighted integral of the sample spectral density matrix is given in the asymptotic framework $p/n \to c \in (0,\infty)$. This result can then be used to develop a non-parametric estimates of certain eigenvalue distributions related to the coefficient matrices. Empirical results showcase the finite-sample performance.

Title: ***Spatial adaptivity and asymptotic quasi exponential invariance***

Speaker: **Moritz Jirak**, TU-Braunschweig, Germany

Schedule: *Monday 11 June, 17:30–18:00*

Abstract: Given a d-dimensional Gaussian observation, it is well-known that finding a non-noise signal in the single components requires the price $\sqrt{\log d}$ in the worst case. In this talk, we analyze the situation if relevant spatial dependence is present. It turns out that the minimax

rate can be expressed in terms of quantiles q_α of the ℓ_∞-norm. Moreover, an optimal, spatially adaptive test can be constructed based on a bootstrap. A key property that allows the construction of this test is an interesting probabilistic feature of the quantiles, a quasi-exponential invariance in their significance level α.

Invited Session IS-12 • Monday 11, 16:30-18:30 • Room Procida

RECENT ADVANCES IN NON-STATIONARY PROCESSES

Organizer(s): Anna Ewa Dudek (AGH University of Science and Technology)

Chair: Anna Ewa Dudek (AGH University of Science and Technology)

Title: ***Minimax optimal detection of structure for multivariate data: application in nonstationary processes analysis.***

Speaker: **Jean-Marc Freyermuth**, Université Aix-Marseille, France

Schedule: *Monday 11 June, 16:30–17:00*

Abstract: We discuss the problem of structure detection in the analysis of nonstationary processes characteristics (alike time varying spectral density). We use the relation between the Sobol decomposition of a multivariate function and the geometry of the hyperbolic wavelet basis to construct test statistics for detecting the presence of some general structure (such as the atomic dimension). We study their performance in the multivariate Gaussian White Noise model with anisotropic estimand and prove that they are minimax optimal.
This is based on joint work with John Aston, Florent Autin, Gerda Claeskens and Christophe Pouet.

Title: ***Time-frequency representation of nonstationary random processes arising from differential equations***

Speaker: **Lorenzo Galleani**, Politecnico di Torino, Italy

Schedule: *Monday 11 June, 17:00–17:30*

Abstract: Random processes from the real world are commonly modeled as the solution of differential equations. Since such processes are inherently nonstationary, their frequency content changes with time, and it can be described by using time-frequency analysis, an ensemble of techniques for the representation of the time-varying spectrum of deterministic and random signals. Time-frequency analysis can be applied directly to the differential equation in the time domain, and the result is a differential equation in time and frequency whose properties clarify the spectral mechanisms involved in the generation

of nonstationary random processes. First, the transformation method is discussed. Then, some examples are illustrated, both for scalar random processes, modeled by a single differential equation, as well as for vector random processes, modeled by systems of differential equations.

Title: ***Absolute regularity of semi-contractive GARCH-type processes***
Speaker: **Michael H. Neumann**, Friedrich-Schiller-Universität Jena, Germany
Co-author(s): Paul Doukhan
Schedule: *Monday 11 June, 17:30–18:00*
Abstract: We prove existence and uniqueness of a stationary distribution and absolute regularity for nonlinear GARCH and INGARCH models of order (p,q). In contrast to previous work we impose, besides a geometric drift condition, only a semi-contractive condition which allows us to include models which would be ruled out by a fully contractive condition. This results in a subgeometric rather than the more usual geometric decay rate of the mixing coefficients. The proofs are heavily based on a coupling of two versions of the processes. The talk is based on joint work with Paul Doukhan (University of Cergy-Pontoise).

Title: ***Estimation of the spatial Loève-spectrum and spatial dual-frequency coherence***
Speaker: **Anna Ewa Dudek**, AGH University of Science and Technology, Poland
Co-author(s): J. Aston, D. Dehay, J-M. Freyermuth, D. Scucs
Schedule: *Monday 11 June, 18:00–18:30*
Abstract: In this talk we develop a novel statistical methodology to study collections of replicated spatially localized EEG time series. We introduce a new class of spatiotemporal processes that are time-harmonizable. Our aim is to estimate nonparametrically the spatial Loève-spectrum (SLS) and the spatial dual-frequency coherence (SDFC) function. The estimate of the SLS is obtained by smoothing dual-frequency cross-periodogram at given spatial locations over the space using bivariate kernels. Finally, the average over the replicates is calculated and the estimate of the SDFC can be easily obtained. We present the asymptotic properties of the proposed estimators. The asymptotic confidence intervals for the SLS and the SDFC are difficult to obtain, because the asymptotic covariance matrices depend on unknown parameters. Thus, we propose a bootstrap approach based on the Circular Block Bootstrap method and show its consistency.

POSTER SESSION

Title: ***A comparison of robust nonlinear regression methods by statistical learning***
Author(s): Jan Kalina, Barbora Peštová
Schedule: Monday 11 June, 18:30–19:30
Abstract: Various estimators for the standard nonlinear regression model are compared with a focus on methods which are robust to outlying measurements in the data. The main contribution is a metalearning study which has the aim to predict the most suitable estimator for a particular data set. Here, various versions of the nonlinear least weighted squares estimator are compared with nonlinear least squares, nonlinear least trimmed squares and a nonlinear regression median, where the last is a special case of nonlinear regression quantiles. The metalearning study is performed over a data base of 24 economic data sets. The nonlinear least weighted squares estimator is able to yield the best result for the most data sets. The metalearning study gives advice how to select appropriate weights for the nonlinear least weighted squares, particularly it reveals tests of normality and heteroscedasticity to play a crucial role in finding suitable weights.

Title: ***Panel Bubble Analysis***
Author(s): Ji Won Shin, Dong Wan Shin
Schedule: Monday 11 June, 18:30–19:30
Abstract: In recent research, it has been found that the presence of bubbles is effective in the detection of bubbles through a recursive method among various methods for detecting bubbles. As a bubble test method through repetitive test, a sup augmented Dickey−Fuller method and a generalized version of the extended augmented Dickey−Fuller method are representative. This method can test the presence of bubbles in one piece of data. Based on the idea that we can not test the bubble across the market, we have extended the recently recursive method to the panel test. The size and power of the panel bubble test method were tested by the Monte Carlo method. The size control was good and the power of the panel GSADF was higher than that of the panel SADF. Therefore, it is confirmed that it can be a good statistic by expanding GASDF, which has been

evaluated as the best bubble test statistic, to panel.

Title: ***Applications of Local Gaussian Correlation in Finance***
Author(s): Bård Støve, Dag Tjøstheim, Håkon Otneim, Geir Drage Berentsen
Schedule: Monday 11 June, 18:30–19:30
Abstract: A number of studies have provided evidence that financial returns exhibit asymmetric dependence, such as increased dependence during bear markets, but there seems to be no agreement as to how such asymmetries should be measured. In this talk we will use the the nonparametric dependence measure called Local Gaussian Correlation to study this asymmetry, and also point at applications like measuring financial contagion. We further provide evidence that the dependence can vary in time. Modelling this time dynamics of the dependence may call for a parametric version of the LGC, and we provide some ideas in this direction.

Title: ***Asymptotic properties of a nonparametric regression estimator for left truncated and associated data***
Author(s): Farida Hamrani, Zohra Guessoum
Schedule: Monday 11 June, 18:30–19:30
Abstract: In this work, we focus on some nonparametric estimation problems. More precisely, considering kernel estimators of the regression function when the interest variable is subject to random left truncation.
We aim at establishing some asymptotic properties of the kernel estimator introduced by Ould Saïd et Lemdani (2006) while taking a dependency framework for the data.
We give the rate of almost sure uniform convergence of the kernel regression estimator when the data are left truncated and associated. Moreover, we show that the suitably standardized estimator is asymptotically normal. We give also illustrations of our results on simulated data.

Title: ***Short-term Estimation of Mobile Traffic by the VAR Model***
Author(s): Jae hyouck Jang
Schedule: Monday 11 June, 18:30–19:30
Abstract: The prediction of mobile traffic is important in many domains, including spectrum requirement, network management and spectrum policy. In South Korea, the volume of mobile traffic reached 315,152 TB at the end of 2017. Most of all, 4G traffic has continuously increased over the years and accounted for 99.5% among several different generations–2G, 3G and 4G. The purpose of this research is to estimate a short-term mobile traffic by using Vector Autoregressive

(VAR) models that are based on monthly multivariate time-series data from January 2012 to December 2017. The estimation was proven to be useful by the mean absolute percentage error (MAPE) through comparing the estimation with the actual result. The relative errors are all less than 3.0%, which show a reasonable prediction of the properties of real mobile traffic. According to the selection of the variables in the model and 2 models were suggested the further prediction for the next 12 months.

Title: ***Extending the cluster-mass test to combined hypothesis***
Author(s): Jaromil Frossard, Olivier Renaud
Schedule: Monday 11 June, 18:30–19:30
Abstract: To analyse data from experiments using EEG, we test a null hypothesis at each time point of the signal $H_0^t: \beta_t = 0$ for $t \in 1,\dots,T$ where β_t is the difference between two experimental conditions at time t, with T usually larger than 400. To take into account the multiple hypotheses, permutation tests using the cluster-mass statistics control the FWER and are a powerful approach when effects happen by cluster. However, usually effects are composed of positive difference immediately followed by a negative difference (or vice versa) between two experimental conditions. Because of zero crossing of the parameters and of the statistics, cluster-mass statistics using F-test or t-test are sub-optimal. We propose to use cluster-mass statistics on the combined null hypothesis that both the mean and the slope of the signal are zero: $H_0^t: \beta_t = 0$ and $H_0^t: \dot{\beta}_t = 0$ for $t \in 1,\dots,T$ where $\dot{\beta}_t$ is the difference in slope between the experimental conditions. Using this combined hypotheses with the cluster-mass statistics, we create bigger clusters which results in more powerful test while keeping the FWER at the nominal level. We compare different methods of smoothing used to produce the estimate of the slope of the signals.

Title: ***Bootstrapping in generalized dynamic factor models***
Author(s): Alexander Braumann, Jens-Peter Kreiss
Schedule: Monday 11 June, 18:30–19:30
Abstract: We consider bootstrap methods for the parameters of singular vector autoregressive models for a static factor process, which is an unobserved part of a generalized dynamic factor model as proposed in Forni et al. 2000, Forni and Lippi 2001, and discussed in Forni et al. 2009, Deistler et al. 2010. 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, which are modelled by a vector

autoregressive system with singular innovation covariance matrix, need to be estimated before estimating the autoregressive parameters. We derive the asymptotic distribution of the autoregressive parameter estimates and analyse under which circumstances these estimates are asymptotically biased and what influences this bias. This is an extension of the univariate case presented in Shintani and Guo 2016. Bootstrap methods are discussed which account for that asymptotic bias. Their asymptotic validity is shown using the asymptotic framework of Bai 2003, Goncalves and Perron 2014 and results from Goncalves and Perron 2016. We compare the finite sample performance of the bootstrap methods and the approach using the asymptotic normal approximation in a simulation exercise.

Title: ***New Second-Order Asymptotic Methods for Nonlinear Models***
Author(s): Gubhinder Kundhi, Paul Rilstone
Schedule: Monday 11 June, 18:30–19:30
Abstract: New higher-order asymptotic methods for nonlinear models are developed. These include generic methods for deriving stochastic expansions of arbitrary order, methods for evaluating the moments of polynomials of sample averages, a method for deriving the approximate moments of the stochastic expansions with simplified expressions for the first four moments and a third-order approximate Saddlepoint expansion. These techniques are applied to improve inferences with the Two-Stage Least Squares estimator and the weak instruments problem. It is well established that Instrumental Variable (IV) estimators in the presence of weak instruments can be poorly behaved, in particular, be quite biased in finite samples. In our application, finite sample approximations to the distributions of the IV estimators are obtained using Edgeworth and Saddlepoint expansions. Higher order analytical corrections provided by these expansions are used to analyze departures from normality. In a Monte-Carlo experiment, the performance of these expansions is compared to the first order approximation and other techniques commonly used in finite samples such as the bootstrap.

Title: ***Some properties of averaged autoregression quantile***
Author(s): Xeniya Yermolenko
Schedule: Monday 11 June, 18:30–19:30
Abstract: Consider the autoregressive model $X_t = \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$. The scalar statistic$B_n(\alpha) = \widehat{\beta}_0(\alpha) + n^{-1} \sum_{t=1}^{n} (\mathbf{y}_{t-1}^*)^\top \widehat{\beta}_1(\alpha)$ is called the averaged autoregression α-quantile. It provides a useful tool in the nonparametric inference on

the autoregressive time series, e.g. in the inference on functionals of distribution of the innovations ε_t, or on the shape of their [unknown] distribution function F. The autors will mainly concentrate on the finite-sample behavior of $B_n(\cdot)$. It better illustrates the structure of the whole problem then the asymptotics, which sometimes involves an excessive simplification. We take recourse to the study of the asymptotic behavior under $n \to \infty$ when the finite sample properties of $B_n(\cdot)$ are getting technically difficult. However, one important asymptotic property we should emphasize: Namely, the average autoregression quantile $B_n(\alpha)$ is getting asymptotically equivalent to the α-quantile of the innovations. This, along with other properties, will be a subject of the further study.

Title: ***Bandwidth selection in kernel distribution estimation for grouped data***

Author(s): Mario Francisco-Fernández, Miguel Reyes, Ricardo Cao, Daniel Barreiro-Ures

Schedule: Monday 11 June, 18:30–19:30

Abstract: Interval-grouped data are defined, in general, when the event of interest cannot be directly observed and it is only known to have been occurred within an interval. In this framework, a nonparametric kernel distribution estimator that is a simple modification of the classical nonparametric kernel distribution estimator is proposed and studied. Some asymptotic properties of this estimator (bias, variance and mean integrated square error) are derived. Additionally, two bandwidth selection methods, plug-in and bootstrap, to be used in this context are proposed and analyzed. The bootstrap method is easy to implement, and does not requiere Monte Carlo. A complete simulation study compares the behavior of the estimator and the bandwidth selectors considering different scenarios of data grouping. The work is completed with an application to real data. In practice, it is observed that the correct behavior of the plug-in bandwidth is limited to a certain degree of grouping. However, the bootstrap bandwidth selector, using an appropriate criterion to select the corresponding pilot bandwidth, presents a stable and unbiased sampling distribution under any scenario or sample size in the simulation studies performed.

Title: ***Non-parametric estimators for estimating bivariate survival function under randomly censored and truncated data***

Author(s): Marialuisa Restaino, Hongsheng Dai, Huan Wang

Schedule: Monday 11 June, 18:30–19:30

Abstract: In bivariate survival analysis it is common to dealt with incomplete

information of the data, due to random censoring and random truncation. Bivariate survival function estimation when both components are censored and truncated has received considerable attention recently. These methods used an iterative computing method which is computationally heavy. Some authors proposed an estimator based on a polar coordinate transformation, which does not require iterative calculations and its large sample properties are established. Starting from their paper, we extend their methods to a class of estimators, based on different data transformations. In particular assuming that the components are both random truncation and random censoring, we propose a class of nonparametric estimators for the bivariate survival function. The proposed class of nonparametric estimators uses such parametric information via a data transformation approach and thus provides more accurate estimates than existing methods without using such information. The large sample properties of the new class of estimators and a general guidance of how to find a good data transformation are given. The proposed method is also justified via a simulation study and an application on an economic data set.

Title: ***Reconstructing missing data sequences in multivariate time series: an application to environmental data***

Author(s): Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna

Schedule: Monday 11 June, 18:30–19:30

Abstract: Missing data arise in many statistical analyses, due to faults in data acquisition, and can have a significant effect on the conclusions that can be drawn from the data. In multivariate time series, moreover, it may happen that not only isolated values but also long sequences of some of the time series' components may miss. In such cases, it is quite impossible to reconstruct the missing sequences basing on the serial dependence structure alone. In this work, we propose an new procedure that aims to reconstruct the missing sequences by exploiting the spatial correlation and the serial correlation of the multivariate time series, simultaneously. The proposed procedure is based on a spatial-dynamic model and imputes the missing values in the time series basing on a linear combination of the neighbor contemporary observations and their lagged values. It is specifically oriented to spatio-temporal data, although it is general enough to be applied to generic stationary multivariate time-series. In this paper, the procedure has been applied to the pollution data, where the problem of missing sequences is of serious concern, with a

remarkably satisfactory performance.

Title: ***On the Asymmetric Impact of Macro–variables on Volatility***
Author(s): Alessandra Amendola, Vincenzo Candila, Giampiero M. Gallo
Schedule: Monday 11 June, 18:30–19:30
Abstract: We extend the GARCH–MIDAS model to take into account possible different impacts from
positive and negative macroeconomic variations on financial market volatility. We evaluate the proposed specification by a Monte Carlo simulation which shows good estimation properties with the increase in the sample size. The empirical application is performed on the daily S&P 500 volatility dynamics with the U.S. monthly industrial production and national activity index as additional (signed) determinants. In the out-of-sample analysis, our proposed GARCH–MIDAS model statistically outperforms the competing specifications, represented by the GARCH, GJR-GARCH and GARCH-MIDAS models.

Invited Session IS-13 • Tuesday 12, 08:30-10:30 • Room Vietri

COMPUTER-INTENSIVE METHODS AND TIME SERIES
Organizer(s): Dimitris Politis (UCSD)
Chair: Dimitris Politis (UCSD)

Title: ***Functional CLT for negatively associated survey sampling and resamplings***
Speaker: **Patrice Bertail**, Université Paris Nanterre, France
Co-author(s): Antoine Rebecq
Schedule: *Tuesday 12 June, 08:30–09:00*
Abstract: For many applications in survey sampling (frequently used for big datasets) or resampling procedures it is important to obtain functional central limit theorems : this allow to obtain asymptotics for empirical processes (yielding for instance uniform condifence intervals for cdf) but also for many functionals having smooth properties (Frechet or Hadamard differentiable in some adequate space). Several partial results have been recently obtained in the framework of survey sampling either by considering particular cases of empirical processes (cf. Boistard et al (2018) or particular type of surveys. In this talk we will show how it is possible to generalize the ideas of Bertail, Chautru and Clemencon (2017), Scand. J. of Stat. proved in the framework of Poisson or conditional Poisson sampling to much more general sampling plans. These results apply on very general ergodic

assumptions on the data generating process.

Title: ***Improved non-parametric inference for the dependence structure of multivariate extremes***
Speaker: **Nan Zou**, University of Toronto
Co-author(s): Axel Bücher, Stanislav Volgushev
Schedule: *Tuesday 12 June, 09:00–09:30*
Abstract: The classic block maxima method aims to estimate the extreme value distribution with maximas over blocks of data. The errors in this estimation are two fold. First, the extreme value distribution of interest characterizes the maxima over an infinitely long block. However, given finite amount of data, one can at most get the distribution of the maxima over a finitely long block. This difference between infinite and finite block length causes a deterministic bias. Second, the distribution of the maxima over a finitely long block is deterministic. Given the randomness of data, one can at most come up with an estimation. This difference between an estimation and the truth gives a stochastic error. In a generic multivariate weakly-dependent setting, a new estimator is proposed to diminish these two errors. To bring down stochastic error, sliding blocks, instead of disjoint blocks, are included. To reduce the deterministic error, a bias term is estimated and then removed. The resulting estimator has been proven to enjoy consistency, asymptotic normality, a variance of a smaller size, and a bias of a smaller order.

Title: ***Bootstrap Based Inference for High Dimensional Vector Autoregressive Models***
Speaker: **Efstathios Paparoditis**, University of Cyprus, Cyprus
Co-author(s): Jonas Krampe, Jens-Peter Kreiss
Schedule: *Tuesday 12 June, 09:30–10:00*
Abstract: Fitting vector autoregressive models to high dimensional time series has attracted considerable interest during the last years. In this talk we will propose appropriate bootstrap methods to infer properties of estimators, to construct individual and simultaneous confidence intervals and to perform individual and group test of parameters. Asymptotic validity of the bootstrap method proposed is established and simulations are presented which demonstrate the finite sample performance of the bootstrap methods.

Title: ***Predictive inference for locally stationary time series with an application to climate data***
Speaker: **Dimitris Politis**, Dept of Mathematics, UCSD, United States of America

Co-author(s): Srinjoy Das
Schedule: *Tuesday 12 June, 10:00–10:30*
Abstract: We show how the Model-free Prediction Principle of Politis (2015) can be applied to handle time series that are only locally stationary, i.e., they can be assumed to be as stationary only over short time-windows. Surprisingly there is little literature on point prediction for general locally stationary time series even in model-based setups and there is no literature on the construction of prediction intervals of locally stationary time series. We attempt to fill this gap here as well. 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. Applications to simulated and real data are included, including analysis of a speleothem dataset that is relevant for the study of climate.

Invited Session IS-14 • Tuesday 12, 08:30-10:30 • Room Ravello

NEW DIRECTIONS IN NONPARAMETRIC METHODS FOR MODERN-DAY PROBLEMS

Organizer(s): Debajyoti Sinha (Florida State University)
Chair: Debajyoti Sinha (Florida State University)

Title: ***Linear Rank tests in complex sample survey data with survival outcomes***
Speaker: **Stuart Lipsitz**, Brigham and Women's Hospital, United States of America
Schedule: *Tuesday 12 June, 08:30–09:00*
Abstract: Linear rank tests (logrank, Wilcoxon, etc) are commonly used to compare survival outcomes across two groups. Unfortunately, no simple extension of the linear rank tests have been proposed for complex sample survey data. With a random sample of independent subjects, any of the linear rank test statistics can be shown to equal a Cox partial likelihood score test, where the particular linear rank test is determined by a weight given to each risk set in the partial likelihood. For example, a weight of 1 gives the log-rank test, and a weight equal to the number at risk gives the Wilcoxon test. For complex survey data, we formulate our extension of the linear rank

tests as an estimating equations score statistic for no group effect in the partial likelihood. The proposed method is applied to an NHANES study which explores if a dash diet (versus a normal diet) reduces the risk of death.

Title: ***Asymptotic for relative frequency when population is driven by arbitrary evolution***
Speaker: **Silvano Fiorin**, Università di Padova, Italy
Schedule: *Tuesday 12 June, 09:00–09:30*
Abstract: Strongly consistent estimates are shown, via relative frequency, for the probability of "white balls" inside a dichotomous urn when such a probability is an arbitrary unknown continuous time dependent function over a bounded time interval. The asymptotic behaviour of relative frequency is studied in a nonstationary context using a Riemann-Dini type theorem for SLLN of random variables with arbitrarily different expectations;furthermore the theoretical results concerning the SLLN can be applied for estimating the mean function of unknown form of a general nonstationary process.

Title: ***Sparse inference with spike-and-slab posterior distributions***
Speaker: **Ismael Castillo**, Sorbonne Université, France
Schedule: *Tuesday 12 June, 09:30–10:00*
Abstract: This talk discusses so-called spike-and-slab prior distributions on unknown sparse vectors. The proportion of non-zero coefficients is chosen using marginal maximum likelihood Empirical Bayes. We consider convergence of the full posterior distribution in terms of rates, its ability to provide asymptotically valid confidence sets that cover the true sparse parameter, as well as the possibility to use such a posterior for multiple testing purposes.

Invited Session IS-15 • Tuesday 12, 08:30-10:30 • Room Furore

HIGH-DIMENSIONAL MATRIX ESTIMATION: THRESHOLDING, SHRINKAGE AND RANDOM MATRIX THEORY

Organizer(s): Matan Gavish (Hebrew University)
Chair: Matan Gavish (Hebrew University)

Title: ***Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising***
Speaker: **Charles Deledalle**, CNRS - Universté de Bordeaux, France
Co-author(s): Jérémie Bigot, Delphine Féral

Schedule: *Tuesday 12 June, 08:30–09:00*

Abstract: We consider the problem of estimating a low-rank signal matrix from noisy measurements under the assumption that the distribution of the data matrix belongs to an exponential family. In this setting, we derive generalized Stein's unbiased risk estimation (SURE) formulas that hold for any spectral estimators which shrink the singular values of the data matrix. This leads to new data-driven spectral estimators, whose optimality is discussed using tools from random matrix theory in the Gaussian spiked population model and through numerical experiments. Our approach also leads to new procedures for singular values shrinkage in finite-dimensional matrix denoising for Gamma-distributed and Poisson-distributed measurements.

Title: ***Prediction in the linearly transformed spiked model***

Speaker: **William Edward Leeb**, Princeton University, United States of America

Schedule: *Tuesday 12 June, 09:00–09:30*

Abstract: The linearly transformed spiked model generalizes the well-known spiked covariance model of Johnstone by incorporating the action of a linear filter or projection on the signal component. As such, it is a more realistic model for many applications, including missing data problems and image deconvolution. In this talk, I will describe the derivation of optimal predictors for this model in high dimensions. I will also explain surprising results contrasting in-sample and out-of-sample prediction, which are new even for the standard spiked model.

Title: ***Robust Sparse Covariance Estimation by thresholding Tyler's M-estimator***

Speaker: **Boaz Nadler**, Weizmann Institute of Science, Israel

Co-author(s): John Goes, Gilad Lerman

Schedule: *Tuesday 12 June, 09:30–10:00*

Abstract: 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, in this work 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 their ratio tending to a positive constant. These bounds are minimax rate-optimal. Results on simulated data support our theoretical analysis.

Title: ***The Noise-Sensitivity Phase Transition in Spectral Group Synchronization Over Compact Groups***
Speaker: **Elad Romanov**, Hebrew University of Jerusalem, Israel
Co-author(s): Matan Gavish
Schedule: *Tuesday 12 June, 10:00–10:30*
Abstract: In Group Synchronization, one attempts to find a collection of unknown group elements from noisy measurements of their pairwise differences. Several important problems in vision and data analysis reduce to group synchronization over various compact groups. Spectral group synchronization is a commonly used, robust algorithm for solving group synchronization problems. It relies on diagonalization of block matrices, whose blocks are matrix representations of the measured pairwise differences. Assuming uniformly distributed measurement errors, we present a rigorous analysis of the accuracy and noise sensitivity of spectral group synchronization algorithms over a general compact group. We find the large-sample asymptotic expected error of spectral group synchronization, as well as the asymptotically exact location of a critical noise level, beyond which it necessarily fails. Our work generalizes and fills the gaps in A. Singer's original analysis of spectral group synchronization.
Joint work with Matan Gavish.

Invited Session IS-16 • Tuesday 12, 08:30-10:30 • Room Positano

MEASUREMENT ERROR PROBLEMS

Organizer(s): Ingrid Van Keilegom (KU Leuven)
Chair: Ingrid Van Keilegom (KU Leuven)

Title: ***Conditional density estimation with covariate measurement error***
Speaker: **Xianzheng Huang**, University of South Carolina, United States of America
Co-author(s): Haiming Zhou
Schedule: *Tuesday 12 June, 08:30–09:00*
Abstract: We consider estimating the density of a response conditioning on an error-prone covariate. Motivated by two existing kernel density estimators in the absence of covariate measurement error, we propose a method to correct the existing estimators for measurement error. We investigate asymptotic properties of the resultant estimators under different types of measurement error distributions. Moreover, we

adjust bandwidths readily available from existing bandwidth selection methods developed for error-free data to obtain bandwidths for the new estimators. Extensive simulation studies are carried out to compare the proposed estimators with naive estimators that ignore measurement error, which also provide empirical evidence for the effectiveness of the proposed bandwidth selection methods. A real-life data example is used to illustrate implementation of these methods under practical scenarios.

Title: ***Estimation of Boundary of a Variable observed with Symmetric Error***

Speaker: **Jean-Pierre Florens**, Toulouse School of Economics, France

Co-author(s): Leopold Simar, Ingrid Van Keilegom

Schedule: *Tuesday 12 June, 09:00–09:30*

Abstract: Consider the model $Y = X + \varepsilon$ with $X = \tau + Z$, where τ is an unknown constant (the boundary of X), Z is a random variable defined on $\Re^+$, ε is a symmetric error, and ε and Z are independent. Based on a iid sample of Y we aim at identifying and estimating the boundary τ when the law of ε is unknown (apart from symmetry) and in particular its variance is unknown. We propose an estimation procedure based on a minimal distance approach and by making use of Laguerre polynomials. Asymptotic results as well as finite sample simulations are shown. The paper also proposes an extension to stochastic frontier analysis, where the model is conditional to observed variables. The model becomes $Y = \tau(w_1, w_2) + Z + \varepsilon$, where Y is a cost, w_1 are the observed outputs and w_2 represents the observed values of other conditioning variables, so Z is the cost inefficiency. Some simulations illustrate again how the approach works in finite samples.

Title: ***On relaxing the distributional assumption of stochastic frontier models***

Speaker: **Hohsuk Noh**, Sookmyung Women's University, Korea, Republic of (South Korea)

Co-author(s): Ingrid Van Keilegom

Schedule: *Tuesday 12 June, 09:30–10:00*

Abstract: The stochastic frontier model has been considered as an alternative of deterministic frontier models in that it allows both measurement error and inefficiency. However, such merit is often dimmed by strong assumption about measurement error and inefficiency such as normal-half normal pair or normal-exponential pair. Since the distribution of the measurement error has almost universally accepted as being normal in many situations, here we show how to estimate various stochastic frontier models with a relaxed assumption about

the inefficiency distribution relying on the recent work of Kneip and his coworkers. We also provide some theoretical results about the estimation.

Title: ***Nonparametric Adjustment for Measurement Error in Time to Event Data: Application to Risk Prediction Models***
Speaker: **Malka Gorfine**, Tel Aviv University, Israel
Schedule: *Tuesday 12 June, 10:00–10:30*
Abstract: Mismeasured time to event data used as a predictor in risk prediction models will lead to inaccurate predictions. This arises in the context of self-reported family history, a time to event predictor often measured with error, used in Mendelian risk prediction models. Using validation data, we propose a method to adjust for this type of error. We estimate the measurement error process using a nonparametric smoothed Kaplan-Meier estimator, and use Monte Carlo integration to implement the adjustment. We apply our method to simulated data in the context of both Mendelian and multivariate survival prediction models. Simulations are evaluated using measures of mean squared error of prediction (MSEP), area under the response operating characteristics curve (ROC-AUC), and the ratio of observed to expected number of events. These results show that our method mitigates the effects of measurement error mainly by improving calibration and total accuracy. We illustrate our method in the context of Mendelian risk prediction models focusing on misreporting of breast cancer, fitting the measurement error model on data from the University of California at Irvine, and applying our method to counselees from the Cancer Genetics Network. We show that our method improves overall calibration, especially in low risk deciles.

Invited Session IS-17 • Tuesday 12, 08:30-10:30 • Room Amalfi

FUNCTIONAL DATA ANALYSIS I
Organizer(s): Alicia Nieto Reyes (Universidad de Cantabria)
Chair: Alicia Nieto Reyes (Universidad de Cantabria)

Title: ***A depth-based method for functional time series forecasting with dynamic updating***
Speaker: **Raúl Jiménez**, Universidad Carlos III de Madrid, Spain
Co-author(s): Antonio Elias
Schedule: *Tuesday 12 June, 08:30–09:00*
Abstract: An approach is presented for making predictions about functional

time series with dynamic updating. These are functional data obtained by slicing an almost continuous time record into natural consecutive periods, when the last of them has not yet been fully observed. The method is applied to data coming from periodically correlated processes, electricity demand, and other case studies. Both for the case studies and simulations, we provide accurate point forecast and tight bands that preserve the shape of the unobserved curve segment, covering it on high proportion for a given confidence level. The approach is computationally efficient, completely data-driven, and substantially different to time series methods, which are based on statistical models of temporal correlation.

Title: ***Domain selection for functional data: Extending false discovery rate to p-values defined on continuous domains***

Speaker: **Simone Vantini**, MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy.

Co-author(s): Niels Lundtorp Olsen, Alessia Pini

Schedule: *Tuesday 12 June, 09:00–09:30*

Abstract: A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of the multiple comparisons problem. We define and discuss the notion of false discovery rate (FDR) in the setting of functional data defined on a compact set. Moreover, a continuous version of the Benjamini-Hochberg procedure is introduced along with a definition of adjusted p-value function. Some general conditions are stated, under which the continuous Benjamini-Hochberg procedure provides control of FDR. The proposed method is applied to satellite measurements of Earth temperature. In detail, we aim at identifying the regions of the planet where temperature has significantly increased in the last decades. For sake of comparison, we also introduce a spherical analogue of the interval-wise testing procedure (i.e., a recent non-parametric tool for local inference) able to continuously control the family-wise error rate on spherical caps. Due to the different scopes and theoretical properties of the methods, there are notable differences in the areas that are selected, with FDR being less conservative.

Title: ***On the Optimal Reconstruction of Partially Observed Functional Data***

Speaker: **Dominik Liebl**, Universität Bonn, Germany

Co-author(s): Alois Kneip

Schedule: *Tuesday 12 June, 09:30–10:00*

Abstract: We propose a new reconstruction operator that 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. We show the optimality of our reconstruction operator and demonstrate that the usually considered regression operators generally cannot be optimal reconstruction operators. Our estimation theory allows for autocorrelated functional data and considers the practically relevant situation in which each of the n functions is observed at m discretization points. We derive rates of consistency for our nonparametric estimation procedures using a double asymptotic ($n \to \infty, m \to \infty$). For data situations, as in our real data application where m is considerably smaller than n, we show that our functional principal components based estimator can provide better rates of convergence than any conventional nonparametric smoothing method.

Title: ***Estimating a covariance function from fragments of functional data***

Speaker: **Aurore Delaigle**, University of Melbourne, Australia

Schedule: *Tuesday 12 June, 10:00–10:30*

Abstract: Functional data are often observed only partially, in the form of fragments. In that case, the standard approaches for estimating the covariance function do not work because entire parts of the domain are completely unobserved. In previous work, Delaigle and Hall (2013, 2016) have suggested ways of estimating the covariance function, based for example on Markov assumptions. In this work we take a completely different approach which does not rely on such assumptions. We show that, using a tensor product approach, it is possible to reconstruct the covariance function using observations located only on the diagonal of its domain.

Invited Session IS-18 • Tuesday 12, 08:30-10:30 • Room Procida

LIKELIHOOD-FREE METHODS OF INFERENCE

Organizer(s): Laura Ventura (University of Padova)

Chair: Laura Ventura (University of Padova)

Title: ***A Robust Bayesian Exponentially Tilted Empirical Likelihood Method***

Speaker: **Catherine Forbes**, Monash University, Australia

Co-author(s): Zhichao Liu, Heather Anderson

Schedule: *Tuesday 12 June, 08:30–09:00*

Abstract: This paper proposes a new Bayesian approach for analysing moment condition models using data that may be contaminated by 'outliers'. Building on the Bayesian exponentially tilted empirical likelihood (BETEL) approach of Schennach (2005), auxiliary variables are used in conjunction with relevant moment conditions to stochastically trim potential outliers from the desired posterior distribution. We also demonstrate that both the BETEL and the new robust BETEL (RBETEL) posteriors may be linked to the recent work of Bissiri, Holmes and Walker (2016) who propose a general framework for updating prior belief via a specified loss function. In addition to an empirical illustration, the results of simulation experiments will be reviewed.

Title: ***Mean and median bias-reducing estimating equations for meta analysis***

Speaker: **Ioannis Kosmidis**, University of Warwick, United Kingdom

Co-author(s): Annamaria Guolo, Sophia Kyriakou, Nicola Sartori, Cristiano Varin

Schedule: *Tuesday 12 June, 09:00–09:30*

Abstract: This talk focuses on recent work on bias-reducing estimating equations for the heterogeneity parameter in meta-analysis and meta-regression settings. These estimating equations result in estimators that have mean or median bias of lower order than the maximum likelihood one. We will discuss how estimation can be performed via a convenient two-step coordinate descent process, and how the estimating equations give rise to new pivots with enhanced inference performance for meta-analytic hypotheses. The new pivots are also found to be robust under model misspecification. The exposition of the methodology will be accompanied by illustrations with real-data case studies.

Title: ***Saddlepoint approximations for short and long memory time series: a frequency domain approach***

Speaker: **Elvezio M. Ronchetti**, University of Geneva, Switzerland

Co-author(s): Davide La Vecchia

Schedule: *Tuesday 12 June, 09:30–10:00*

Abstract: Saddlepoint techniques provide numerically accurate, small sample approximations to the distribution of estimators and test statistics. Except for a few simple models, these approximations are not available in the framework of stationary time series. We contribute to fill this gap. Under short or long range serial dependence, for Gaussian and non Gaussian processes, we show how to derive and implement saddlepoint approximations for two relevant classes of

frequency domain statistics: ratio statistics and Whittle's estimator. We compare our new approximations to the ones obtained by the standard asymptotic theory and by two widely-applied bootstrap methods. The numerical exercises for Whittle's estimator show that our approximations yield accuracy's improvements, while preserving analytical tractability. A real data example concludes the paper.

Title: ***Approximate Bayesian analysis with adjusted composite likelihoods with an application to meta-analysis with binary outcomes***
Speaker: **Nicola Sartori**, University of Padova, Italy
Co-author(s): Michele Lambardi di San Miniato
Schedule: *Tuesday 12 June, 10:00–10:30*
Abstract: We explore the use of composite likelihoods for Bayesian inference in under-specified models, in which, for instance, some marginal distributions are assumed, but higher-order dependences in the data are ignored. Calibration and adjustment methods have been proposed in the literature to allow the use of composite likelihoods in Bayesian analysis. We propose a new simple adjustment method that can be used in studies in which the focus is on specific parameters of interest. The comparison of different adjustments, under a given prior, is performed using a simulation-based procedure that allows to validate the use of the composite likelihood in the Bayesian framework. Results are illustrated in a model for meta-analysis with binary outcomes.

Plenary Session PL-02 • Tuesday 12, 11:00-12:00 • Auditorium Tafuri

ASYMPTOTIC BEHAVIOUR OF CREDIBLE REGIONS – THE ADAPTIVE CASE

by Judith Rousseau
University of Oxford, UK

Chair: Brunero Liseo (Sapienza Università di Roma)

Abstract: Bayesian credible regions are typical Bayesian measures of uncertainty and are a natural by product of the Bayesian approach, which makes it a driving force of Bayesian approaches. By construction they depend in particular on the choice of the prior distribution, wich in high or infinite dimensional can be very influential. To shed light on the impact of the prior model in complex or high dimensional models, we study the asymptotic frequentist behaviour of the posterior distri-

bution in general and of some credible regions in particular.
In this talk I will present some of the results that have been obtained on credible balls in non-parametric models and in high dimensional models using prior models that have a discrete hierarchical structure. For these priors we show that in a rather general framework inflated credible balls have good frequentist coverage uniformly over a large range of parameter values, at the cost of a $\sqrt{\log(n)}$ inflation of the radius. We will also show that, at least in some cases, this inflation is necessary and we will discuss the implications of this result.

Contributed Session CS-09 • Tuesday 12, 12:00-13:00 • Room Ravello

NETWORK DATA

Chair: Swati Chandna (Birkbeck, University of London)

Title: ***Quantifying Uncertainty in Node Feature Mining of Large Social Networks***

Speaker: **Vyacheslav Lyubchich**, University of Maryland Center for Environmental Science, United States of America

Co-author(s): Cuneyt G. Akcora, Yulia R. Gel, Murat Kantarcioglu

Schedule: *Tuesday 12 June, 12:00–12:20*

Abstract: Continuously growing online social networks raise the importance of new scalable and reliable algorithms for estimating node features, such as individual's socio-demographics, preferences, or health status. In this project, we pioneer the research on uncertainty quantification and uncertainty propagation in network feature mining. In particular, we evaluate the uncertainty of network-based estimators and use data-driven techniques to define the sample size required for different levels of confidence. To accomplish that, we propose a novel bootstrap method for uncertainty analysis of node features in social network mining, derive its asymptotic properties, and demonstrate its effectiveness with extensive experiments. Furthermore, we propose to implement a metric of dispersion of the estimates to assist in selecting the optimal sampling parameters in an automatic and data-driven way.

Title: ***Time Series Modeling on Dynamic Networks***

Speaker: **Jonas Krampe**, TU Braunschweig, Germany

Schedule: *Tuesday 12 June, 12:20–12:40*

Abstract: We consider multivariate time series on dynamic networks with a fixed number of vertices. Each component of the time series is

assigned to a vertex of the underlying network. The dependency of the various components of the time series is modeled dynamically by means of the edges.
We make use of a multivariate doubly stochastic time series framework that is we assume linear processes for which the coefficient matrices are stochastic processes themselves. We explicitly allow for dependence in the dynamics of the coefficient matrices, including of course an i.i.d. structure as is typically assumed in random coefficients models. In the paper asymptotic normality of simple statistics like the sample mean is investigated. Furthermore, autoregressive moving average models are defined in this framework. Estimators of network autoregressive models are discussed and how this can be used to forecast such a process. Some interesting features of these processes are shown in simulations and the finite sample behavior of the forecast approach is investigated.

Title: ***Nonparametric regression with network data***
Speaker: **Swati Chandna**, Birkbeck, University of London, United Kingdom
Co-author(s): Pierre-André Maugis
Schedule: *Tuesday 12 June, 12:40–13:00*
Abstract: We introduce nonparametric methods addressing the setting where a sample of small networks, along with additional information, is observed. For example, in a connectome study, for each individual in the sample both a structural brain network is observed, along with covariates such as age, gender, etc. We work under the framework of exchangeability commonly used to model network data where the node labels carry no information. Under this formulation, estimation of the limit object termed 'graphon', has attracted significant attention in the nonparametric literature on networks. Building upon the standard graphon model, we provide a framework that can test for any given node presenting significantly different behavior across different values of the covariates. Further, we find that although a significant portion of the graphon literature focuses on block-model approximations of the graphon, in our setting full nonparametric inference is possible and computationally tractable. We illustrate our approach using a set of brain network observations from multiple individuals.

Contributed Session CS-10 • Tuesday 12, 12:00-13:00 • Room Furore

SMOOTHING TECHNIQUES I

Chair: Laura Borrajo (Universidade da Coruña)

Title: ***Nonparametric Estimation of Time-Varying Parameters in Nonlinear Models***

Speaker: **Dennis Kristensen**, University College London, United Kingdom

Co-author(s): Young Jun Lee

Schedule: *Tuesday 12 June, 12:00–12:20*

Abstract: We propose nonparametric estimators of time-varying parameters in a general class of non-linear time series models. Under weak regularity conditions, we show the proposed estimators are consistent and follow a normal distribution in large samples. A key concept in our analysis is local stationarity for which we provide primitive conditions to hold in the case of Markov processes. To demonstrate the usefulness of our general results, we provide primitive conditions for our theory to apply in a number of examples, including ARCH models and Poisson autoregressions with time-varying parameters.

Title: ***Smooth Nonparametric Survival Analysis***

Speaker: **Dimitris Avraam Ioannides**, University of Macedonia, Greece

Schedule: *Tuesday 12 June, 12:20–12:40*

Abstract: We present three methodological advances in the area of nonparametric survival analysis under random right censoring. The first development is the construction of pointwise confidence intervals for the survival function by means of a central limit theorem for an, already existing in the literature, kernel smooth survival estimate. The second development is a central limit theorem for the Integrated Square Error which is subsequently applied on developing a goodness of fit test. Establishment of its asymptotic distribution yields the proposed test statistic for drawing decision on the null hypothesis of correctness of the assumed survival function. Numerical simulations quantify the empirical powers of the suggested test for certain confidence levels. Finally, we prove that the relative deficiency of the Kaplan–Meier estimate with respect to a specific kernel survival function tends to infinity as the sample size goes to infinity. The last result advocates the use in practice of the kernel based estimate over the Kaplan–Meier when performance is measured through the mean square error.

Title: ***Nonparametric probability mass function estimation for big-but-biased data***

Speaker: **Laura Borrajo**, Universidade da Coruña (Spain); CITIC (Spain)
Co-author(s): Ricardo Cao, Sofia Olhede, Swati Chandna
Schedule: *Tuesday 12 June, 12:40–13:00*
Abstract: It is often argued that in big data setups "numbers speak for themselves". However, some authors have recently warned about the validity of this idea due to the common presence of sampling bias in big data. In fact, several problems coming from ignoring this bias have been recently reported. A fully nonparametric approach is considered in this work. The probability mass function estimation problem is studied in this setup over categorical data, in a nonparametric framework, when the biasing weight function is known (unrealistic) as well as for unknown weight functions (realistic). In addition to the big-but-biased sample, a small sized simple random sample of the real population is considered. An estimator involving both samples is proposed to remedy the problem of ignoring the weight function. Asymptotic expressions for the mean squared error of the estimator proposed are considered. This leads to some asymptotic formulas for the optimal smoothing parameters. Simulated datasets are used to illustrate the performance of the estimator proposed in this work. This method is also applied to the study of a dataset related to food allergies.

Contributed Session CS-11 • Tuesday 12, 12:00-13:00 • Room Positano

ADVANCES IN PERMUTATION TESTING

Chair: Stefano Bonnini (University of Ferrara)

Title: ***Nonparametric Permutation-based Inference on Round-Robin Design for Clustered Data***
Speaker: **Livio Corain**, Department of Management and Engineering, University of Padova, Italy
Co-author(s): Rosa Arboretti, RIccardo Ceccato, Luigi Salmaso
Schedule: *Tuesday 12 June, 12:00–12:20*
Abstract: Round-robin design refers to settings where data are originated from joining or interacting between two units so that from each unit an outcome arises in the form of two dependent responses. Sometimes the dependency among pairs of observations is driven by an hierarchical factor, i.e. the units are clustered within a nested factor. In case of multivariate responses and by assuming a two-level fixed effect data representation model, the goal of the present

paper is to propose an extension of the nonparametric permutation and combination-based methodology suitable for testing on both within- and between-cluster effects of round-robin clustered data design. Moreover, by extending a recently proposed methodology aimed at ranking on multivariate populations, results of testing are then exploited to provide a ranking of cluster effects. Via a Monte-Carlo simulation study we investigated the properties of the proposed testing and ranking methodology and we proved its validity under different random distributions. Finally, by using play-by-play basketball data we present an application to player-based data sport analytics.

Title: ***From Paper to Program: Challenges of Implementing Permutation Tests***

Speaker: **Kellie Nicole Ottoboni**, UC Berkeley, United States of America

Schedule: *Tuesday 12 June, 12:20–12:40*

Abstract: As the field of data science grows and computational resources abound, it is increasingly common for people to analyze their data using permutation tests. Introductory statistics textbooks often teach parametric statistics exclusively; when they do teach permutation tests, they are often portrayed as being model-free. In fact, permutation tests do require assumptions, and this approach to teaching omits crucial considerations. Without appropriate pedagogical material about issues of implementation, people are more likely to misuse permutation tests. To fill this gap, we are developing a short open access textbook. Topics from the book include: how controlling for covariates can improve statistical power, illustrated using clinical trial data with multiple time points and study sites; how to design a permutation test for a complex experiment, illustrated with a study to assess gender bias in teaching evaluations; and computational limitations of sampling algorithms and pseudorandom number generators for big data. These examples are written in accessible language, shown using real-world data, and accompanied by R and Python code. Our materials should supplement, not replace, instruction of permutation test theory, and will enable intermediate statistics students, data scientists, and domain researchers to design the best nonparametric tests for their observational or experimental studies.

Title: ***Statistical Inference in Behavioral Economics: a Permutation Approach***

Speaker: **Stefano Bonnini**, University of Ferrara, Italy

Co-author(s): Isabel María Parra Oller

Schedule: *Tuesday 12 June, 12:40–13:00*

Abstract: Some experiments of Behavioral Economics concern problems of Intertemporal Choice (IC). People tends to prefer immediate gains and to postpone losses. The discount rate is the proportion or percentage of increase in value needed to compensate a 1-year delay. In the classic theory of Discounting Utility Model, it is constant with respect to time delay, amount and sign (gain or loss). Behavioral anomalies such as magnitude effect (discount rate decreasing function of the amount), sign effect (losses discounted less than gains) and delay effect (discount rate higher for shorter time horizon) can be tested with complex experiments. In these experiments, groups of people are interviewed about their intertemporal choices to study individual and group discount rates and test for anomalies in their IC. From the inferential point of view, suitable advanced testing methods are needed in order to face the complexity of the hypotheses under test. For example multivariate and multistrata testing methods for repeated measures with directional alternatives are necessary. The high number of variables makes the problem complex and requires advanced solutions. We propose a non-parametric method based on a permutation approach, which has many advantages and good properties, proved through Monte Carlo simulation studies.

Contributed Session CS-12 • Tuesday 12, 12:00-13:00 • Room Procida

IMPERFECTLY OBSERVED DATA I

Chair: Samuel Maistre (Université de Strasbourg)

Title: ***Analysis of extreme values for censored data under competing risks***

Speaker: **Julien Worms**, Université de Versailles-St-Quentin, France

Co-author(s): Rym Worms

Schedule: *Tuesday 12 June, 12:00–12:20*

Abstract: When randomly censored values of a variable X are observed, one is often interested in values which are expectations $\theta = E(\phi(X))$ of a functional of X : expectation of X or particular values $F(x)$ of the distribution function are examples of such situations. Some literature exists concerning the conditions on ϕ which guarantee that θ can be correctly estimated by the integral of ϕ with respect to the Kaplan-Meier distribution F_n. However, when the function ϕ is unbounded, depends on the sample size n, and does not vanish in the neighborhood of the upper endpoint, no general result exists.

An example of such a situation will be described in this talk : if we are interested in the extreme values of X, one can construct an estimator of the extreme value index of F which is a Kaplan-Meier integral of a function ϕ_n which is unbounded and never vanishes in the neighborhood of the upper endpoint of F. With such an estimator, extreme quantiles of F can be estimated. Focus will be on heavy tailed distributions, but we will show how the approach can be extended to light tails or presence of competing risks (with the Aalen-Johansen estimator replacing the Kaplan-Meier one).

Title: ***Limit distribution of Kaplan-Meier double integrals***
Speaker: **Nicolas Rivera**, University of Cambridge, United Kingdom
Co-author(s): Tamara Fernandez
Schedule: *Tuesday 12 June, 12:20–12:40*
Abstract: We study asymptotic properties of the Kaplan-Meier U-statistic Φ_n, defined as $\Phi_n = \int_{x \neq y} \phi(x,y) d\hat{F}_n(x) d\hat{F}_n(y)$, where $\hat{F}_n$ is the Kaplan-Meier estimator for the distribution function under independent right censoring, and ϕ is a symmetric function. In particular, we prove that under appropriate standardisation the Kaplan-Meier U-statistic Φ_n has asymptotically Normal distribution, while in a degenerated case its asymptotic distribution is the same as a (potentially infinity) linear combination of independent χ^2 random variables. As a corollary of our proof, for the degenerated case, we obtain an asymptotic representation of Φ as a (standard) U-statistic of independent random variables. Our proof is based on extending well-known martingale techniques, developed by Gill, to the context of bivariate integrals, and a bivariate version of the Advance and Backward operators defined by Ritov and Wellner. The statement of our results and proof techniques are clearer than previous approaches, and, at the same time, our conditions are much weaker and almost optimal.

Title: ***Conditional estimators under random censorship on the whole real line***
Speaker: **Samuel Maistre**, Université de Strasbourg, France
Co-author(s): Ingrid Van Keilegom
Schedule: *Tuesday 12 June, 12:40–13:00*
Abstract: Either truncation or strong independence conditions are always assumed when one uses conditional estimators under random right censoring, e.g. the Beran estimator for conditional cumulative distribution function estimation. A result on the conditional cumulative hazard rate is presented, avoiding truncation. The unconditional literature, which uses martingale theory, is adjusted to the conditional case.

Special Invited Session SIS-03 • Tuesday 12, 12:00-13:00 • Auditorium Tafuri

NONPARAMETRIC FIRST-ORDER ANALYSIS OF SPATIAL AND SPATIO-TEMPORAL POINT PROCESSES WITH APPLICATIONS.

by Wenceslao González Manteiga[(1)] **& María Isabel Borrajo**[(1)] **& Isabel Fuentes-Santos**[(2)]

[(1)] University of Santiago de Compostela, Spain; [(2)] Spanish National Research Council, Spain

Speaker: Wenceslao González Manteiga
Chair: Michael G. Schimek (Medical University of Graz)

Abstract: In the point process framework, kernel intensity estimation has been limited to exploratory analysis due to its lack of consistency. This work addresses different procedures to obtain a consistent estimator of the first order intensity such as kernel estimation of the density of event locations and kernel intensity estimation. We propose a smooth bootstrap procedure for inhomogeneous point processes (including the case with information of covariates) in order to develop effective bandwidth selectors for kernel intensity estimation. The consistent estimators introduced above, are used to estimate the first order intensity of the wildfires registered in Galicia during a long period of years. Finally this kind of estimators are used for two problems of interest:

a) The nonparametric comparison of first-order intensity functions and

b) One separability test for spatio-temporal point process.

Invited Session IS-19 • Tuesday 12, 14:30-16:30 • Room Vietri

NONPARAMETRICS FOR NETWORKS, EVENT COUNTING AND FUNCTIONAL DATA

Organizer(s): Enno Mammen (Heidelberg University)
Chair: Enno Mammen (Heidelberg University)

Title: ***Nonparametric Approaches for Exponential Random Graph Models***

Speaker: **Göran Kauermann**, LMU Muenchen, Germany
Schedule: *Tuesday 12 June, 14:30–15:00*
Abstract: The Exponential Random Graph Model (ERGM) is the central tool in statistical network data analysis. The model allows for intuitive interpretations but suffer from so called degeneracy problems when the network becomes large. The models behave peculiar in large networks with thousands of actors. Standard models containing two-star or triangle counts as statistics are often unstable leading to completely full or empty networks. Moreover, numerical methods break down which makes it complicated to apply. These problems, to some extend, trace from the overall linear structure of the models. We propose to replace linear statistics by nonparametric terms, which itself are fit using penalized spline smoothing. This approach extends the model class of ERGMs and links statistical network data analysis to nonparametric statistics. To be specific we propose two strategies to circumvent common obstacles in ERGMs for large networks. First, we use a subsampling scheme to obtain (conditionally) independent observations for model fitting and secondly, we show how linear statistics (like two-stars etc.) can be replaced by smooth functional components. These two steps in combination allow to fit stable models to large network data, which is illustrated by an example with data from facebook

Title: ***Using dynamic dependence for non-parametric inference for continuous-time event counting and link-based dynamic network models***
Speaker: **Alexander Kreiß**, Heidelberg University, Germany
Co-author(s): Enno Mammen, Wolfgang Polonik
Schedule: *Tuesday 12 June, 15:00–15:30*
Abstract: A flexible modelling approach based on counting processes for both dynamic event counting (in a network of actors) and dynamic link-based networks is studied. The main interest lies in non-parametric likelihood based inference of a parameter function depending on time via kernel smoothing. A rigorous analysis of the asymptotic behaviour of the estimates is carried out by allowing the number of actors to tend to infinity. The main difficulty lies in modelling the dependence among the actors. It is allowed that the dependence structure among the actors is random and changes over the observation period. In the modelling this is respected by mixing and conditional independence concepts which take these dynamics into account. The finite sample performance of the estimators is illustrated through an empirical analysis of bike share data.

Title: ***Additive Regression with Hilbertian Responses***

Speaker: **Byeong Park**, Seoul national University, Korea, Republic of (South Korea)

Co-author(s): Jeong Min Jeon

Schedule: *Tuesday 12 June, 15:30–16:00*

Abstract: This paper develops a foundation of methodology and theory for the estimation of structured nonparametric regression models with Hilbertian responses.

Our method and theory are focused on the additive model, while the main ideas may be adapted to other structured models. For this, the notion of Bochner integration is introduced for Banach-space-valued maps as a generalization of Lebesgue integration.

Several statistical properties of Bochner integrals, relevant for our method and theory, and also of importance in their own right, are presented for the first time. Our theory is complete. The existence of our estimators and the convergence of a practical algorithm that evaluates the estimators are established. These results are non-asymptotic as well as asymptotic. Furthermore, it is proved that the estimators achieve the univariate rates in pointwise, L^2 and uniform convergence, and that the estimators of the component maps converge jointly in distribution to Gaussian random elements. Our numerical examples include the cases of functional, density-valued and simplex-valued responses, which demonstrates the validity of our approach.

Title: ***Modeling and Prediction of Dynamic Networks using Binary Autoregressive Time Series Processes***

Speaker: **Carsten Jentsch**, TU Dortmund, Germany

Co-author(s): Shaikh Tanvir Hossain, Lena Reichmann

Schedule: *Tuesday 12 June, 16:00–16:30*

Abstract: Suppose a time series of networks is identified by their adjacency matrices $A_1, \ldots, A_T$, where $A_t = (a_{ij;t})_{i,j=1,\ldots,N}$ with $a_{ij;t} \in \{0,1\}$ and $a_{ij,t} = 1$ indicating that there is a directed edge pointing from vertex i to vertex j at time t. To model the joint dynamics of the edges, we propose to use multivariate binary time series processes. For this purpose, we adopt the class of Discrete AutoRegressive Moving-Average (DARMA) models introduced for univariate categorical data by Jacobs and Lewis (1983). Recent extensions of these models allow the application to vector-valued data and to model negative autocorrelations by a simple modification. The resulting model class is flexible enough to capture very general autocorrelations driving the dynamic network structure. For the purely autoregressive case, Yule-Walker-type equations hold that allow in principle an explicit

estimation of all model parameters. However, as the dimension of the adjacency matrices grows quadratically with the number of vertices, we shall make use of Lasso-penalization techniques to estimate sparse models. We adopt the approach of Basu and Michailidis (2015), who established consistent estimation for high-dimensional VAR models under sparsity. Our modeling approach is suitable for prediction of edge probabilities in dynamic networks. We illustrate our method by simulations and for real data.

Invited Session IS-20 • Tuesday 12, 14:30-16:30 • Room Ravello

STATISTICAL INFERENCE FOR COMPLEX DATA STRUCTURES

Organizer(s): Soumendra Lahiri (NCSU)

Chair: Francesco Giordano (University of Salerno)

Title: ***Fast bootstrap strategies for massive data***

Speaker: **Srijan Sengupta**, Virginia Tech, United States of America

Schedule: *Tuesday 12 June, 14:30–15:00*

Abstract: The bootstrap is a popular and powerful method for statistical inference in a remarkable variety of scenarios. However, for massive datasets which are increasingly prevalent, the bootstrap becomes prohibitively costly in computation and its feasibility is questionable even with modern parallel computing platforms. It is therefore important to come up with computationally efficient alternatives of bootstrap that retain the flexibility and statistical properties of classical bootstrap, while being computationally scalable. This talk will cover a suite of recently developed fast bootstrap strategies for independent as well as dependent data — the Subsampled Double Bootstrap (SDB), the Bag of Little Bootstraps (BLB), the Aggregation of Little Bootstraps (ALB), and the Subsampled Residual Bootstrap (SRB).

Title: ***Joint circular distributions in view of higher order spectra of time series***

Speaker: **Masanobu Taniguchi**, Waseda University, Japan

Co-author(s): Shogo Kato, Hiroaki Ogata, Arthur Pewsey

Schedule: *Tuesday 12 June, 15:00–15:30*

Abstract: Circular data analysis is emerging as an important component of statistics. For this half century, various circular distributions have been proposed, e.g., von Mises distribution, wrapped Cauchy distribution, among other things. In this talk we introduce a class of

joint circular distributions from the higher order spectra of time series, which can describe very general joint circular distributions. Hence we can develop the statistical inference for dependent observations on the circle. We present a family of distributions on the circle derived from the ARMA spectral density. It is seen that the proposed family includes some existing circular families as special cases. For these special cases, the normalizing constant and trigonometric moments are shown to have simple and closed form. We develop the asymptotic optimal inference theory based on the local asymptotic normality (LAN) on the circle. Because the observations are permitted to be dependent, the theory opens a new paradigm in the estimation for joint circular distributions.

Title: ***A Resampling Test of Symmetry in 3-D Rotations***
Speaker: **Ulrike Genschel**, Iowa State University, United States of America
Schedule: *Tuesday 12 June, 15:30–16:00*
Abstract: Three-dimensional orientation data arise in various scientific studies, such as human kinematics, structural geology, and materials science. With such data, each observation is represented by a 3x3 rotation to denote the orientation of some object after rotating its reference frame away from some "central" reference frame in 3-d. In many applications, it is of interest to investigate whether a random sample of orientations has a symmetric distribution, whereby observations can be interpreted as directionally symmetric random perturbations of an underling mean-location rotation parameter. Many common models for random rotations assume such distributional symmetry, but an approach to formally assess this feature is lacking for orientation data. We provide a general characterization of distributional symmetry for random rotations, using an angle-axis representation of 3x3 rotations. Under the assumption of symmetry, a random rotation is induced by three independent random variables, with two variables having known (uniform-type) distributions. From this, we develop a test statistic for distributional symmetry and investigate a convenient bootstrap procedure for approximating the complex sampling distribution of this statistic.

Title: ***Varying coefficient models for causal inference: estimation and testing with smooth backfitting***
Speaker: **Stefan Sperlich**, University of Geneva, Switzerland
Schedule: *Tuesday 12 June, 16:00–16:30*
Abstract: In causal inference one often looks at average treatment effects on mean or quantile functions.
Today, popular approaches to reflect the heterogeneity of those effects

are either looking at the entire distribution of the target variable or works with regression models containing random coefficients. In contrast to these standard approached, working with generalized structured models like generalized additive or varying coefficient models allows the empirical researcher to explicitly model the heterogeneity of impacts. These can ease a lot the interpretation and consequently also the deriving of recommendations e.g. for policy making.
We first highlight the advantages of this method compared to others and then present estimators and test procedure based on smoothed backfitting. Simulation studies show excellent performance, and some data examples the usefulness of this approach. The presentation is based on different papers with different co-authors.

Invited Session IS-21 • Tuesday 12, 14:30-16:30 • Room Furore

BAYES IN HIGH DIMENSION FROM FREQUENTIST PERSPECTIVE

Organizer(s): Subhashis Ghoshal (North Carolina State University)
Chair: Subhashis Ghoshal (North Carolina State University)

Title: *epsilon*-***Approximations to the Pitman-Yor process***
Speaker: **Pierpaolo De Blasi**, University of Torino, Italy
Schedule: *Tuesday 12 June, 14:30–15:00*
Abstract: We consider approximations to the Pitman-Yor process obtained by truncating the stick-breaking representation. The truncation level is defined by a random stopping rule as the minimum number of frequencies that sum up to a probability larger than 1-*epsilon*. The construction allows to control the approximation error in total variation distance with probability one. We derive the limiting distribution of the truncation level as the approximation error epsilon goes to zero in terms of a polynomially tilted positive stable distribution. The usefulness of this theoretical result is demonstrated by devising a sampling algorithm to approximate functionals of the epsilon-version of the Pitman-Yor process.

Title: ***The Bayes Lepski's Method and Credible Bands through Volume of Tubular Neighborhoods***
Speaker: **William Weimin Yoo**, Leiden University, Netherlands, The
Co-author(s): Aad W. van der Vaart
Schedule: *Tuesday 12 June, 15:00–15:30*
Abstract: For a general class of priors based on random series basis expansion,

we develop the Bayes Lepski's method to estimate unknown regression function. In this approach, the series truncation point is determined based on a stopping rule that balances the posterior mean bias and the posterior standard deviation. Equipped with this mechanism, we present a method to construct adaptive Bayesian credible bands, where this statistical task is reformulated into a problem in geometry, and the band's radius is computed based on finding the volume of certain tubular neighborhood embedded on a unit sphere. We consider two special cases involving B-splines and wavelets, and discuss some interesting consequences such as the uncertainty principle and self-similarity. Lastly, we show how to program the Bayes Lepski stopping rule on a computer, and numerical simulations in conjunction with our theoretical investigations concur that this is a promising Bayesian uncertainty quantification procedure.

Title: ***Empirical Bayes uncertainty quantification for regression***
Speaker: **Eduard Belitser**, VU Amsterdam
Co-author(s): Subhashis Ghosal
Schedule: *Tuesday 12 June, 15:30–16:00*
Abstract: We propose an empirical Bayes method for high dimensional linear regression models. Following an oracle approach which quantifies the error locally for each possible value of the parameter, we show that the empirical Bayes posterior contracts at the optimal rate at all parameters and leads to uniform size-optimal credible balls with guaranteed coverage under an "excessive bias restriction" condition. This condition gives rise to a new slicing of the entire space that is suitable for ensuring uniformity in uncertainty quantification. The obtained results immediately lead to optimal contraction and coverage properties for many conceivable classes simultaneously. The results are also extended to high dimensional additive nonparametric regression models.

Title: ***Bayesian inference in nonparanormal graphical models***
Speaker: **Subhashis Ghoshal**, North Carolina State University, United States of America
Schedule: *Tuesday 12 June, 16:00–16:30*
Abstract: Gaussian graphical models, where it is assumed that the variables of interest jointly follow multivariate normal distributions with sparse precision matrices, have been used to study intrinsic dependence among several variables, but the Gaussianity assumption may be restrictive in many applications. A nonparanormal graphical model is a nonparametric generalization of a Gaussian graphical model for continuous variables where it is assumed that the variables

follow a Gaussian graphical model only after some unknown smooth monotone transformation. Priors on the underlying transformations are put using finite random series of B-splines with increasing coefficients. On the underlying precision matrix of the transformed variables, we consider either a spike-and-slab prior or a continuous shrinkage prior on its Cholesky decomposition. A Hamiltonian Monte-Carlo method allows efficient sampling from the posterior distribution of the transformation. We show that the posterior distribution for the transformation function is consistent under general conditions. We study the numerical performance of the proposed method through a simulation study and apply it on a real dataset.
Based on a joint work with Jami J. Mulgrave.

Invited Session IS-22 • Tuesday 12, 14:30-16:30 • Room Positano

MODERN NONPARAMETRIC AND RANKING METHODS

Organizer(s): Philip Yu (The University of Hong Kong)
Chair: Philip Yu (The University of Hong Kong)

Title: ***Change of Measure Applications in Nonparametric Statistics***
Speaker: **Mayer Alvo**, University of Ottawa, Canada
Schedule: *Tuesday 12 June, 14:30–15:00*
Abstract: Neyman (1937) was the first to propose a change in measure in the context of goodness of fit problems. This provided an alternative density to the one for the null hypothesis. Hoeffding introduced a change of measure formula for the ranks of the observed data which led to obtaining locally most powerful rank tests. In this paper, we review these methods and propose a new approach which leads on the one hand to new derivations of existing statistics. On the other hand, we exploit these methods to obtain Bayesian applications for ranking data.

Title: ***Graphical linear models for paired comparisons data***
Speaker: **Ori Davidov**, University of Haifa, Israel
Schedule: *Tuesday 12 June, 15:00–15:30*
Abstract: Graph based least square procedures for rank data are analyzed from a statistical perspective. Finite sample properties as well as asymptotic distribution theory is developed. Goodness of fit tests for detecting local and global lack of fit are discussed. The methodology is explored by simulation. A real data example is analyzed.

Title: ***Distance-based and distance-free rank models: some new developments***

Speaker: **Michael G. Schimek**, Medical University of Graz, Austria

Schedule: *Tuesday 12 June, 15:30–16:00*

Abstract: The Mallows rank model is one of the most popular approaches to consensus ranking when a fixed set of items is ranked by different assessors. For the estimation of the latent parameters a distance function between the rankings is required. For most distance measures exact inference is highly demanding because the partition function necessary for the normalisation of the model is very expensive to compute. Most literature concerns the Mollows model with Kendall distance because then the partition function can be computed analytically. Maximum likelihood estimation is prohibitive (NP-hard), therefore various heuristic algorithms were proposed. An alternative approach has been introduced by Vitelli et al. (2018, JMLR, forthcoming): a Bayesian Mallows model that works with any right-invariant distance. Estimation is obtained via a Metropolis-Hastings iterative algorithm. For the exact partition function it converges to the Bayesian posterior distribution. Another recent proposal replaces the distance-based Mallows model by a simple auxiliary model in the context of indirect inference (Svendova and Schimek; 2017, CSDA 115, 122-135). The latent parameters are evaluated by means of a Metropolis-Hastings McMC algorithm. Unlike Bayesian approaches, the quantification of uncertainty requires in addition bootstraps from the empirical ranks. Both types of rank models are discussed and compared.

Title: ***Weighted distance-based models for ranking data using the R package rankdist***

Speaker: **Philip Yu**, The University of Hong Kong, Hong Kong S.A.R. (China)

Co-author(s): Zhaozhi Qian

Schedule: *Tuesday 12 June, 16:00–16:30*

Abstract: rankdist is a recently developed R package which implements various distance-based ranking models. These models capture the occurring probability of rankings based on the distances between them. The package provides a framework for fitting and evaluating finite mixture of distance-based models. This paper also presents a new probability model for ranking data based on a new notion of weighted Kendall distance. The new model is flexible and more interpretable than the existing models. We show that the new model has an analytic form of the probability mass function and the maximum likelihood estimates of the model parameters can be obtained efficiently even for ranking

involving a large number of objects.

Invited Session IS-23 • Tuesday 12, 14:30-16:30 • Room Maiori

GOODNESS-OF-FIT METHODS AND RELATED PROBLEMS I

Organizer(s): Simos Meintanis (National and Kapodistrian University of Athens)

Chair: Simos Meintanis (National and Kapodistrian University of Athens)

Title: ***Nonparametric estimation of the cross ratio function with applications***

Speaker: **Jan Willem Hendrik Swanepoel**, North-West University, Potchefstroom, South Africa,

Co-author(s): Paul Janssen, Noël Veraverbeke, Steven Abrams

Schedule: *Tuesday 12 June, 14:30–15:00*

Abstract: The cross ratio function is a commonly used tool to describe local dependence between correlated event times. Being a ratio of conditional hazard functions, the cross ratio can be rewritten in terms of the survival copula of these correlated failure times. Bernstein estimators for this survival copula are used to define a nonparametric estimator of the cross ratio and its asymptotic normality is derived. Using attractive computational formulas for the survival copula, a simulation study is conducted to determine the finite sample performance of the estimator for copulas with different types of local dependency. A dataset on asthma attacks in children is used to investigate the local dependence between event times in the placebo and treated group. The estimated cross ratio reveals that the anti-allergic drug disconnects the local dependence between the time to the first and time to the second attack.

Title: ***Robustness-efficiency trade-offs in minimum distance estimation of the mixed Poisson-lognormal model***

Speaker: **Tomasz Hołyński**, Vienna University of Technology, Vienna, Austria

Co-author(s): Jacek Haneczok

Schedule: *Tuesday 12 June, 15:00–15:30*

Abstract: This study concerns the minimum distance estimation in the two-parameter mixed Poisson-lognormal distribution. Despite applicability of that model in diverse fields, such as biostatistics, insurance, computational linguistics and operations research, a robust and efficient estimation remains a challenging task (limited in literature to the unreliable moment matching and unrobust likelihood

maximization). We conduct broad simulation experiments of estimators based on various distances (divergence measures), yielding robust and relatively efficient alternatives. Our main focus is on estimators based on empirical probability generating function (epgf). We consider the single-point-squared and the integrated-squared distances between the epgf and its theoretical counterpart. Performance evaluation is carried out in terms of joint mean squared errors as well as their ranks, in presence and absence of excessive outliers. Our assessment indicates a superior performance of the epgf-based estimators under contamination and their minor deficiencies in pure-model data (outperforming, for example, the celebrated Hellinger distance estimator). The simulations are accompanied by empirical and analytical influence considerations. Since the Poisson-lognormal distribution has no closed form, we address also the related computational difficulties. We introduce a new approximation based on the Laplace transform that facilitates accurate computations near to zero and in the far tail.

Title: ***Goodness of fit tests for Pareto and Log-normal distributions***
Speaker: **Emanuele Taufer**, University of Trento, Italy
Co-author(s): Flavio Santi, Giuseppe Espa, Maria Michela Dickson
Schedule: *Tuesday 12 June, 15:30–16:00*
Abstract: The Zenga (1984) inequality curves $\lambda(p)$ and $Z(p)$ have some interesting properties. The former is constant in p for Type I Pareto distributions, while the latter is constant in p for Log-Normal distributions. After discussing in detail these aspects, these characterizing behaviors will be exploited to obtain graphical and analytical tools for tail analysis, estimation and goodness of fit tests. Order statistics-based estimators of the curves will be presented and discussed; furthermore, a testing procedure for Pareto-type behavior and one for Log- normality based on a regression of $\lambda(p)$ and $Z(p)$ against p will be introduced. The properties of the proposed estimation and testing strategies are analyzed theoretically and by means of simulations; comparisons with competing testing strategies are presented. An application to data sets on city sizes, facing the debated issue of distinguishing Pareto-type tails from Log-normal tails, illustrates how the proposed method works in practice.

Title: ***Tests of Independence for Circular Data***
Speaker: **Simos Meintanis**, National and Kapodistrian University of Athens, Athens, Greece; Unit for Business Mathematics and Informatics, North West University, South Africa
Schedule: *Tuesday 12 June, 16:00–16:30*

Abstract: Tests of independence are proposed for a given pair of circular random variables. The tests are based on the empirical trigonometric moments of arbitrary order, and hence are consistent against any given alternative. The methods are also generalized to higher dimension and to the case of testing independence between a random variable distributed over the (hyper)sphere and another variable taking values in the Euclidean space of general dimension.

Invited Session IS-24 • Tuesday 12, 14:30-16:30 • Room Amalfi

NONPARAMETRIC ANALYSIS OF CLUSTERS OF EXTREMES AND ITS APPLICATION

Organizer(s): Natalia Markovich (V.A.Trapeznikov Institute of Control Sciences)

Chair: Natalia Markovich (V.A.Trapeznikov Institute of Control Sciences)

Title: ***On discrimination between two classes of distribution tails***

Speaker: **Igor V. Rodionov**, Lomonosov Moscow State University, Russian Federation; Moscow Institute of Physics and Technology, Russian Federation

Schedule: *Tuesday 12 June, 14:30–15:00*

Abstract: A consistent test of discrimination between two classes of distribution tails using higher order statistics is proposed. We do not assume belonging the corresponding distribution functions to a maximum domain of attraction

Title: ***On the tail index estimation in the block approach***

Speaker: **Marijus Vaičiulis**, Vilnius University, Lithuania

Schedule: *Tuesday 12 June, 15:00–15:30*

Abstract: We consider a tail index estimation when the data can be divided into several blocks but only a few of largest observations within blocks are available for analysis. In particular, we investigate the asymptotic behavior of the parameterized statistic which construction is based on generalized scaled log-spacings of extreme order. Assuming that the distribution of observations satisfies semi-parametric conditions, we show that this statistic is weakly consistent and asymptotically normal. We apply obtained results for construction of the parameterized tail index estimators.

Title: ***Robust Estimation of the Extremal Index***

Speaker: **Maria Ivette Leal de Carvalho Gomes**, CEAUL and FCiencias.ID, Universidade de Lisboa, Portugal

Co-author(s): Cristina Miranda, Manuela Souto de Miranda

Schedule: *Tuesday 12 June, 15:30–16:00*

Abstract: Many examples in the most diverse fields of application show the need for statistical methods of analysis of extremes of dependent data. And a crucial issue that appears when there is dependency is the reliable estimation of the *extremal index* (EI), a parameter related to the clustering of large events. The most popular EI-estimators, like the blocks's EI-estimators, are very sensitive to anomalous cluster sizes. Hence the need for a robust version of such EI-estimator, the main topic under discussion. The proposed robust procedure, developed under the assumption that the distribution of the number of exceedances belongs to some neighborhood of the Poisson model, preserves asymptotic properties and outperforms the blocks estimator in all simulation studies, even without considering contaminated models. Comparisons between blocks and sliding blocks versions show the existence of bias reduction and a consequent improvement on the stability of the estimates, without an increase in the mean square error. Different generalized means have been recently used in a successful estimation of the extreme value index, among other parameters of univariate extreme events, and will be now proposed for the EI-estimation.

Title: ***Nonparametric estimation of extremal index with threshold selection by discrepancy method***

Speaker: **Natalia Markovich**, V.A.Trapeznikov Institute of Control Sciences, Russian Federation

Schedule: *Tuesday 12 June, 16:00–16:30*

Abstract: Nonparametric estimation of extremal index of stochastic processes is considered.

Nonparametric blocks, runs and intervals estimators of the extremal index require the selection of a threshold u. To estimate u we develop the discrepancy method. The latter was proposed first for a nonparametric estimation of probability density functions. The von Mises-Smirnov (M-S) and the Kolmogorov-Smirnov (K-S) statistics were used as the discrepancy measures. Quantiles of limit distributions of these statistics were used as the discrepancy value δ. The method was modified by the author for heavy-tailed densities and the extremal index. To this end, the discrepancy statistics M-S and K-S were calculated not by entire sample but by K largest order statistics. The selection of K and δ is still an open problem.

To overcome this problem we obtain now the limit distribution of the modified M-S statistic regarding the value K. This allows us to select δ using quantiles of the latter distribution. To this aim, we use the exponential limit distribution of the normalized inter-cluster size. The cluster means the number of consecutive observations exceeding threshold u between two consecutive nonexceedances. The exposition is accompanying by simulated examples.

Invited Session IS-25 • Tuesday 12, 14:30-16:30 • Room Procida

NEW DIRECTIONS FOR NONPARAMETRIC METHODS IN DATA SCIENCE

Organizer(s): Tanya Garcia (Texas A&M University)

Chair: Ingrid Van Keilegom (KU Leuven)

Title: ***Semi-supervised Approach to Causal Inference with EHR Data***

Speaker: **Tianxi Cai**, Harvard University, United States of America

Co-author(s): David Chen

Schedule: *Tuesday 12 June, 14:30–15:00*

Abstract: While clinical trials remain a critical source for studying disease risk, progression and treatment response, they have limitations including the generalizability of the study findings to the real world and the limited ability to test broader hypotheses. In recent years, due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with specimen bio-repositories, large integrated EHR datasets now exist as a new source for translational research. These datasets open new opportunities for deriving real-word evidence for causal treatment effects. Yet, they also bring methodological challenges. For example, obtaining validated phenotype information, such as presence of a disease condition and treatment response, is a major bottleneck in EHR research, as it requires laborious medical record review. In this talk, I'll discuss a semi-supervised approach to estimating causal treatment effects when the outcome of interest is only observed in a small subset of patients. These methods will be illustrated using EHR data from Partner's Healthcare.

Title: ***The Hodges-Lehmann estimator in a location mixture model: is asymptotic normality good enough?***

Speaker: **Fadoua Balabdaoui**, ETH Zurich, Switzerland

Schedule: *Tuesday 12 June, 15:00–15:30*

Abstract: We derive the exact limit distribution of the Hogdes-Lehmann estimator of Hunter et al. (2007) which they considered in the semi-

parametric model of a location mixture of symmetric distributions. We give sufficient conditions on the true symmetric component for the weak convergence to hold. As already expected by Hunter et al. (2007), the limit distribution is that of a three-dimensional centered Gaussian distribution. The variance-covariance matrix can be calculated using the known covariance structure of a standard Brownian Bridge. The examples we used to illustrate the theory indicate that the estimator is not to be advocated when the mixture components are not well separated.

Title: ***Using a Surrogate Marker for Early Testing of a Treatment Effect***
Speaker: **Layla Parast**, RAND, United States of America
Co-author(s): Tianxi Cai, Lu Tian
Schedule: *Tuesday 12 June, 15:30–16:00*
Abstract: The development of methods to identify, validate and use surrogate markers to test for a treatment effect has been an area of intense research interest given the potential for valid surrogate markers to reduce the required costs and follow-up times of future studies. Several quantities and procedures have been proposed to assess the utility of a surrogate marker. However, few methods have been proposed to address how one might use the surrogate marker information to test for a treatment effect at an earlier time point, especially in settings where the primary outcome and the surrogate marker are subject to censoring. In this presentation, we propose a novel test statistic to test for a treatment effect using surrogate marker information measured prior to the end of the study in a time-to-event outcome setting. We propose a robust nonparametric estimation procedure and propose inference procedures. In addition, we evaluate the power for the design of a future study based on surrogate marker information. We illustrate the proposed procedure and relative power of the proposed test compared to a test performed at the end of the study using simulation studies and an application to data from the Diabetes Prevention Program.

Title: ***Bayesian functional predictor for disease onset***
Speaker: **Tanya Garcia**, Texas A&M University, United States of America
Schedule: *Tuesday 12 June, 16:00–16:30*
Abstract: Neuroimaging data now plays an imperative role in understanding the progression of neurodegenerative diseases such as Huntington's, Parkinson's and Alzheimer's disease. A primary focus with such data is discovering and evaluating those neural regions most predictive of clinical outcomes such as age of disease onset. Identifying these neural regions is of high public health relevance as it aids to determine

when and how therapeutic treatments should intervene. But handling the complex structure of high-dimensional neuroimaging data, extracting relevant regions and overcoming large censoring rates in event responses complicates image analysis. Overcoming these challenges requires developing advanced statistical tools for neuroimaging analysis which is the focus of this talk. We propose a new Bayesian framework for identifying brain regions of interest associated with disease onset. Our methods are shown to be consistent, and leads to new scientific discoveries in a study of Huntington's disease.

Contributed Session CS-13 • Tuesday 12, 17:00-18:00 • Room Ravello

NONPARAMETRIC INFERENCE AND ESTIMATION II

Chair: Hanan Emad Ahmed (Tilburg University)

Title: ***A ROC curve comparison test with covariates***
Speaker: **Arís Fanjul-Hevia**, Universidade de Santiago de Compostela, Spain
Co-author(s): Wenceslao González Manteiga, Juan Carlos Pardo-Fernández
Schedule: *Tuesday 12 June, 17:00–17:20*
Abstract: The comparison of Receiver Operating Characteristic (ROC) curves is a commonly accepted way to compare the discriminatory capability of certain diagnostic methods. Several procedures may be found in the literature concerning this problem. Along with the diagnostic variables it is usual to observe other covariates, but that extra information has been hardly ever considered for the comparison of this kind of curves. This information can be taken into account when constructing the ROC curves, usually by means of the conditional ROC curve. It is known that the discriminatory capability of these curves may be influenced by this extra information, so it should be included in the study. With this idea, a new non-parametric test is proposed for the comparison of the conditional ROC curves. A bootstrap mechanism is used to calibrate the critical values of the test. Simulations are run to analyse the practical performance of the test in terms of level approximation and power. An application to real data is also presented to illustrate the procedure.

Title: ***About Identification and Estimation in Semi- and Nonparametric Transformation Models***
Speaker: **Nick Kloodt**, Universität Hamburg, Germany
Schedule: *Tuesday 12 June, 17:20–17:40*

Abstract: Over the last years, transformation models have attracted more and more attention since they are often used to satisfy desirable assumptions 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. After a brief motivation for transformation models, some identification results and estimation techniques in semi- and nonparametric transformation models are presented, before the case of a fully nonparametric model with heteroscedastic error terms is considered. Not only an identification result is given, but also estimation techniques for the corresponding transformation function together with some convergence results for this (consistent) estimator are provided.

Title: ***Improved estimation of the extreme value index using related variables***

Speaker: **Hanan Ahmed**, Tilburg University, Netherlands

Co-author(s): John H.J. Einmahl

Schedule: *Tuesday 12 June, 17:40–18:00*

Abstract: Heavy tailed phenomena are naturally analyzed by extreme value theory. A fundamental step in such an analysis is the estimation of the extreme value index (EVI), which describes the tail heaviness of the underlying distribution. In the present paper we consider the situation where we have next to the usual n observations another $n+m$ observations of one or more related variables, like, e.g., hurricane losses and related weather recordings for a longer period than that of the losses. Based on such a data set, we present an adapted version of the Hill estimator that shows greatly improved behavior and establish its asymptotic normality. For this adaptation the tail dependence between the variable of interest and the related variables plays a crucial role. A comprehensive simulation study confirms the substantially improved performance of our adapted estimator relative to the Hill estimator.

Contributed Session CS-14 • Tuesday 12, 17:00-18:00 • Room Furore

STATISTICAL LEARNING

Chair: Denis Larocque (HEC Montreal)

Title: ***Clustering-oriented selection of the amount of smoothing in kernel density estimation***

Speaker: **Alessandro Casa**, Dipartimento di Scienze Statistiche, Università degli Studi di Padova

Co-author(s): Josè Chacón, Giovanna Menardi

Schedule: *Tuesday 12 June, 17:00–17:20*

Abstract: The nonparametric approach to cluster analysis draws a correspondence between groups and the modes of the density assumed to underlie the data. Operationally the identification of the clusters requires the estimation of the density, usually obtained resorting to kernel estimators. When considering these estimators the selection of a parameter controlling the amount of smoothing is needed; this selection constitutes a critical issue since a bad choice could imply an undersmoothed or oversmoothed density estimate, potentially leading to the appearance of spurious modes or to cover interesting features, respectively. A plethora of different bandwidth selectors has been proposed, all aiming to obtain an appropriate density estimate in a global sense. Nonetheless clustering constitutes a different problem with respect to density estimation. The aim is to identify correctly the domain of attractions of the density modes and it could be achieved even using sub-optimal, from a traditional global measure perspective, estimate of the density itself. Hence a criterion to choose the smoothing parameter, specifically tailored for cluster analysis, is proposed and compared with some other well established criteria.

Title: ***Nonparametric PU learning for Bayesian classification***

Speaker: **Vasily Vasilyev**, Moscow Institute of Physics and Technology (State University), Russian Federation

Schedule: *Tuesday 12 June, 17:20–17:40*

Abstract: There are four main types of learning in ML: supervised, semi-supervised, unsupervised and reinforcement learning. The new nonparametric method of binary classification for PU learning is proposed, which is a subdivision of semi-supervised learning. Abbreviation PU means that for purposes of studying analytical model there are two training sets: positive set (consists of objects of zero class) and unlabeled set (includes elements of both classes without label). In this method Bayesian approach with multivariate kernel density estimators and novel method of defining a prior probability of zero class are applied. This method may be employed to independent r.v. as well as dependent processes (e.g. hidden Markov models). The novel method's performance on simulated and real data is compared to the same performance of the optimal Bayesian solution with known pdf (in the case of simulated data) and prior probabilities of both classes. Comparison to the state of

art approaches is also considered.

Title: ***Prediction Intervals Based on Random Forests***
Speaker: **Denis Larocque**, HEC Montreal, Canada
Co-author(s): Marie-Hélène Roy
Schedule: *Tuesday 12 June, 17:40–18:00*
Abstract: The classical and most commonly used approach to building prediction intervals is the parametric approach. However, its main drawback is that its validity and performance highly depend on the assumed functional link between the covariates and the response. This research investigates new and computationally efficient methods that improve the performance of prediction intervals with random forests. Two aspects are explored: The splitting rule and the method used to build the prediction interval. In addition to the default least squares splitting rule, two alternative splitting criteria are investigated. We also present and evaluate the performance of five flexible methods for constructing prediction intervals. This yields 15 distinct method combinations. To reliably attain the desired confidence level, we include a calibration procedure performed on the out-of-bag information provided by the forest. The 15 method combinations are thoroughly investigated, and compared to four alternative methods through simulation studies and in real data settings. The results show that the proposed methods are very competitive. They outperform commonly used methods in both in simulation settings and with real data.

Contributed Session CS-15 • Tuesday 12, 17:00-18:00 • Room Positano

STATISTICAL ALGORITHMS FOR BIG DATA
Chair: Matus Maciak (Charles University)

Title: ***Learning minimum volume sets from regenerative Markov chains***
Speaker: **Gabriela Beata Ciolek**, Telecom ParisTech, France
Co-author(s): Stephan Clemencon, Patrice Bertail
Schedule: *Tuesday 12 June, 17:00–17:20*
Abstract: The main goal of this talk is to present generalization bounds for minimum volume set estimation for regenerative Markov chains. Minimum volume sets can be used to detect anomalies/outliers, determine highest posterior density, multivariate confidence regions or clustering.We introduce a new maximal concentration inequality in order to show that learning rate bounds depend not only on the

complexity of the class of candidate sets but also on the ergodicity rate of the chain X, expressed in terms of tail conditions for the length of the regenerative cycles. Finally, we show that it is straightforward to extend the preceding results to the Harris recurrent case.

Title: ***Model and algorithm for text simplification***
Speaker: **El Mehdi Issouani**, Université Paris-Nanterre, France
Schedule: *Tuesday 12 June, 17:20–17:40*
Abstract: This contribution consists of developing and implementing portable and feasible methods for modeling and automating text simplification. The problem is motivated by automatic transcription of websites to simplified version for deaf persons. Finding effective ways to adapt a tagger which was trained as a first step «with similar linguistic properties» is potentially becoming the standard way of making other tasks such as Parsing, Retrieving, Text summarization & finally Text Simplification. The key to our 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). This allows us to tag large corpus of texts. Parsing complex structure and aligning it with a simplified one may also be seen as a classification and a scoring problem. We will also discuss other methods based on deep neural networks.

Title: ***Quantile LASSO in Nonparametric Models with Changepoints Under Optional Shape Constraints***
Speaker: **Matus Maciak**, Charles University, Czech Republic
Schedule: *Tuesday 12 June, 17:40–18:00*
Abstract: Nonparametric models are popular modeling tools because of their overall flexibility. In our approach, we apply nonparametric techniques for panel data structures with changepoints and optional shape constraints and the estimation is performed in a fully data driven manner by utilizing atomic pursuit methods - LASSO regularization techniques in particular. However, in order to obtain robust estimates and, also, to have a more complex insight into the underlying data structure, we adopt the quantile lasso approach: the estimation process and the following inference become both more challenging but the results are more useful in practical applications. The main theoretical results are presented and some inferential tools for changepoint significance and shape constraints are proposed. The presented methodology is applied for a real data scenario and some finite sample properties are investigated via s simulation study.

Special Invited Session SIS-04 • Tuesday 12, 17:00-18:00 • Auditorium Tafuri

FUNCTIONAL BOXPLOTS FOR MULTIVARIATE CURVES

by Marc G. Genton
King Abdullah University of Science and Technology, Saudi Arabia

Speaker: Marc G. Genton
Chair: Efstathia Bura (TU Wien)

Abstract: A two-stage functional boxplot is introduced for the visualization and exploratory data analysis of multivariate curves. Specifically, the original functional boxplot is combined with an outlier-detection procedure based on the functional directional outlyingness, which accounts for both the magnitude and shape outlyingness of functional data. This combination is robust to various types of outliers and, hence, captures the data structures more accurately than the functional boxplot alone. It also allows for both marginal and joint analysis of the multivariate curves. We apply the proposed tool to various functional datasets.

Invited Session IS-26 • Tuesday 12, 18:00-19:30 • Room Vietri

HIGH-DIMENSIONAL AUTOCOVARIANCE MATRIX ESTIMATION

Organizer(s): Tommaso Proietti (University of Rome Tor Vergata)
Chair: Tommaso Proietti (University of Rome Tor Vergata)

Title: ***Testing independence for multivariate time series via the auto-distance correlation matrix***
Speaker: **Konstantinos Fokianos**, University of Cyprus, Cyprus
Schedule: *Tuesday 12 June, 18:00–18:30*

Abstract: We introduce the matrix multivariate auto-distance covariance and correlation functions for time series, discuss their interpretation and develop consistent estimators for practical implementation. We also develop a test for testing the independent and identically distributed hypothesis for multivariate time series data and show that it performs better than the multivariate Ljung–Box test. We discuss computational aspects and present a data example to illustrate the methodology.

Title: ***Estimating MA parameters through factorization of the autocovariance matrix and an MA-sieve bootstrap***

Speaker: **Timothy McMurry**, University of Virginia, United States of America

Co-author(s): Dimitris Politis

Schedule: *Tuesday 12 June, 18:30–19:00*

Abstract: A new method to estimate the moving-average (MA) coefficients of a stationary time series is proposed. The new approach is based on the modified Cholesky factorization of a consistent estimator of the autocovariance matrix. Convergence rates are established, and the new estimates are used in order to implement an MA-type sieve bootstrap. Finite-sample simulations corroborate the good performance of the proposed methodology.

Title: ***High Order Asymptotic Theory of Shrinkage Estimation for General Statistical Models***

Speaker: **Masanobu Taniguchi**, Waseda University, Japan

Co-author(s): Hiroshi Shiraishi, Takashi Yamashita

Schedule: *Tuesday 12 June, 19:00–19:30*

Abstract: In this paper we develop the high order asymptotic theory of shrinkage estimators for general statistical models, which include dependent processes, multivariate models and regression models, i.e., non-i.i.d. models. Introducing a shrinkage estimator of MLE, we compare it with that of MLE by third-order mean squares error (MSE). A sufficient condition for the shrinkage estimator to improve the MLE will be given in a very general fashion. Our model is described as a curved statistical model $p(\cdot;\theta(u))$, where θ is a parameter of larger model and u is a parameter of interest with $\dim u < \dim \theta$. This setting is especially suitable for estimation of portfolio coefficients u based on mean and variance parameters θ. We also mention the advantage of our shrinkage estimators when the dimension of parameter becomes large. Numerical studies are given, and illuminate an interesting feature of the shrinkage estimator.

Invited Session IS-27 • Tuesday 12, 18:00-19:30 • Room Ravello

MINIMAX AND ADAPTIVE INFERENCE

Organizer(s): Alexandre Tsybakov (University of Paris VI)

Chair: Cristina Butucea (ENSAE)

Title: ***Adaptive minimax density estimation.***

Speaker: **Oleg Lepski**, Aix-Marseille Université, France
Schedule: *Tuesday 12 June, 18:00–18:30*
Abstract: We study the problem of adaptive minimax estimation of probability densities with L_p–loss, $1 \leq p < \infty$ over the scale of anisotropic Nikol'skii classes on $\Re^d$. We fully characterize the behavior of the minimax risk for different relationships between regularity parameters and norm indexes in the definitions of the functional class and of the risk. We discover 7 different regimes of the asymptotics of minimax risk, including inconsistency zone. Our pointwise selection rule from the family of kernel estimators leads to optimally-adaptive (in some regimes) or nearly optimally-adaptive estimator (up to logarithmic factor).

Title: ***Nonparametric density estimation from observations with multiplicative measurement errors***
Speaker: **Alexander Goldenshluger**, university of haifa, Israel
Schedule: *Tuesday 12 June, 18:30–19:00*
Abstract: In this paper we study the problem of pointwise density estimation from observations with multiplicative measurement errors. We elucidate the main feature of this problem: the influence of the estimation point on the estimation accuracy. In particular, we show that, depending on whether this point is separated away from zero or not, there are two different regimes in terms of the rates of convergence of the minimax risk. In both regimes we develop kernel–type density estimators and prove upper bounds on their maximal risk over suitable nonparametric classes of densities. We show that the proposed estimators are rate–optimal by establishing matching lower bounds on the minimax risk. Finally we test our estimation procedure on simulated data.

Title: ***Solution of linear ill-posed problems by model selection and aggregation***
Speaker: **Felix Abramovich**, Tel Aviv University, Israel
Co-author(s): Daniela De Canditiis, Marianna Pensky
Schedule: *Tuesday 12 June, 19:00–19:30*
Abstract: We consider a general statistical linear inverse problem, where the solution is represented via a known (possibly overcomplete) dictionary that allows its sparse representation. We propose two different approaches. A model selection estimator selects a single model by minimizing the penalized empirical risk over all possible models. By contrast with direct problems, the penalty depends on the model itself rather than on its size only as for complexity penalties. A Q-aggregate estimator averages over the entire collection

of estimators with properly chosen weights. Under mild conditions on the dictionary, we establish oracle inequalities both with high probability and in expectation for the two estimators. Moreover, for the latter estimator these inequalities are sharp.

Invited Session IS-28 • Tuesday 12, 18:00-19:30 • Room Furore

NONPARAMETRICS IN FINANCE

Organizer(s): Stefan Sperlich (University of Geneva)
Chair: Stefan Sperlich (University of Geneva)

Title: ***Bias reduction in nonparametric estimation in presence of both categorical and continuous data***
Speaker: **Michael Scholz**, University of Graz, Austria
Schedule: *Tuesday 12 June, 18:00–18:30*
Abstract: We present a bias reducing estimator for nonparametric regression which admits both continuous and categorical data. We use the local-polynomial estimation framework and introduce a multiplicatively guided estimator. A simulation study comparing our approach with conventional nonparametric estimators in the presence of mixed data is included as well as an empirical illustration on financial time series data.

Title: ***Nonparametric Specification Testing of Conditional Asset Pricing Models***
Speaker: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-author(s): Francisco Peñaranda, Stefan Sperlich
Schedule: *Tuesday 12 June, 18:30–19:00*
Abstract: This paper presents a new omnibus specification test of conditional asset pricing models. These models provide constraints that conditional moments of returns and pricing factors must satisfy, but most of them do not provide information on the functional form of those conditional moments. Hence, the main interest of this test is that it is not only robust to functional form misspecification of conditional moments, but it also detects any relationship between pricing errors and conditioning variables. This last issue is of crucial interest for power in testing conditional models. Special emphasis is given on practical issues like bias reduction, adaptive bandwidth choice, rather general but simple requirements on the estimates, and finite sample performance, including the resampling approximations.

Title: ***The Cross-Sectional Distribution of Fund Skill Measures***
Speaker: **Olivier Scaillet**, UNIGE, SFI
Co-author(s): Laurent Barras, Patrick Gagliardini
Schedule: *Tuesday 12 June, 19:00–19:30*
Abstract: We develop an econometric methodology to infer the distribution of performance from a large unbalanced panel of individual mutual funds. We investigate net alphas, gross alphas, and value added estimated from linear factor models. We use a simple kernel density estimator, and show its asymptotic normality under increasing cross-sectional and time series dimensions, and vanishing bandwidth. The empirical analysis investigates the distribution of performance on about two thousand US mutual funds from January 1975 to December 2012. We relate the distribution of performance to fund characteristics such as size, turnover, and flows. The analysis reveals an heterogenous impact of these characteristics across funds.

Invited Session IS-29 • Tuesday 12, 18:00-19:30 • Room Positano

GOODNESS-OF-FIT METHODS AND RELATED PROBLEMS II

Organizer(s): Simos Meintanis (National and Kapodistrian University of Athens)

Chair: Simos Meintanis (National and Kapodistrian University of Athens)

Title: ***A characterization of multinormality and corresponding tests of fit***
Speaker: **M. Dolores Jiménez Gamero**, University of Sevilla, Spain
Co-author(s): Norbert Henze, Simos G. Meintanis
Schedule: *Tuesday 12 June, 18:00–18:30*
Abstract: A novel characterization of multivariate normality is provided. It incorporates both the characteristic function and the moment generating function. It is employed to construct a class of affine invariant, consistent and easy-to-use goodness-of-fit tests for normality. The test statistics are suitably weighted L^2-statistics. It is studied their asymptotic behavior both for i.i.d. observations as well as in the context of testing that the innovation distribution of a multivariate GARCH model is Gaussian. It is also studied the finite-sample behavior of the new tests and compared with alternative existing tests

Title: ***Specification testing in nonparametric AR-ARCH models***
Speaker: **Marie Huskova**, Charles University, Czech Republic
Co-author(s): Natalie Neumeyer, Tobias Niebuhr, Leonie Silk

Schedule: *Tuesday 12 June, 18:30–19:00*

Abstract: The talk concerns an autoregressive time series model with conditional heteroscedasticity is considered, where both conditional mean and conditional variance function are modeled nonparametrically. Tests for the model assumption of independence of innovations from past time series values are suggested. Tests based on weighted L2 -distances of empirical characteristic functions are considered as well as a Cramér-von Mises type test. The asymptotic distributions under the null hypothesis of independence are derived and consistency against fixed alternatives is shown. A smooth autoregressive residual bootstrap procedure is suggested and its performance is shown in a simulation study.

Title: ***Tests for validity of semiparametric transformation models***

Speaker: **Charl Pretorius**, Charles University, Prague, Czech Republic

Co-author(s): Marie Hušková, Natalie Neumeyer, Simos G Meintanis

Schedule: *Tuesday 12 June, 19:00–19:30*

Abstract: Consider an observed response Y which, following a certain transformation $\mathscr{Y}_\vartheta := \mathscr{T}_\vartheta(Y)$, can be expressed as a nonparametric, possibly heteroskedastic, regression model referenced by a vector X of regressors. If this transformation model is indeed valid then, conditionally on X, the values of $\mathscr{Y}_\vartheta$ may be viewed as merely being location-scale shifts of the regression errors for some value of the transformation parameter ϑ. We propose tests for the validity of this model and establish the limiting distribution of the test statistics under the null hypothesis and under alternatives. Since the null distribution is complicated we also suggest a certain resampling procedure in order to approximate the critical values of the tests, and subsequently use this type of resampling in a Monte Carlo study of the finite-sample properties of the new tests. In estimating the model we rely on the methods recently proposed in the literature for the aforementioned transformation model. Our tests however deviate from the available tests in that we employ a test involving characteristic functions rather than distribution functions.

Invited Session IS-30 • Wednesday 13, 08:30-10:30 • Room Vietri

RECENT ADVANCES IN NON-STANDARD PROBLEMS INCLUDING SHAPE-CONSTRAINED INFERENCE AND MIXTURE MODELS

Organizer(s): Moulinath Banerjee (University of Michigan)

Chair: Moulinath Banerjee (University of Michigan)

Title: ***Central limit theorems for the L_p-loss of smooth isotonic estimators***
Speaker: **Rik Lopuhaä**, Delft University of Technology, Netherlands, The
Co-author(s): Eni Musta
Schedule: *Wednesday 13 June, 08:30–09:00*
Abstract: Shape constrained nonparametric estimation dates back in the 1950s. In his milestone paper in 1956, Grenander derived the maximum likelihood (ML) estimator of a nonincreasing density, whereas Brunk (1958) obtained the least squares estimator of a monotone regression function. The last decades, similar estimators have been proposed in other statistical models, including Cox's proportional hazards model with a monotone baseline hazard. Typically, these isotonic estimators are step functions that exhibit a non normal limit distribution at cube-root n rate. On the other hand, a long stream of research has shown that, if one is willing to assume more regularity on the function of interest, smooth estimators can be used to achieve a faster rate of convergence to a Gaussian distributional law and to estimate derivatives. The asymptotic behavior of global distances between the estimator and the function of interest, such as L_p distances and supremum distance, has been studied for traditional isotonic estimators. In this talk, similar results are presented for smooth isotonic estimators.

Title: ***Divide and Conquer methods in monotone regression***
Speaker: **Cécile Durot**, Université Paris Nanterre, France
Co-author(s): Moulinath Banerjee, Bodhisattva Sen
Schedule: *Wednesday 13 June, 09:00–09:30*
Abstract: The divide and conquer principle in studied in the isotonic regression problem, where rates of convergence are slower than the square-root of the sample size, and limit distributions are non-Gaussian. For a fixed model, the pooled estimator obtained by averaging non-standard estimates across mutually exclusive subsamples, outperforms the non-standard monotonicity-constrained (global) estimator based on the entire sample in the sense of point wise estimation. However, this gain in efficiency under a fixed model comes at a price: the pooled estimator's performance, in a uniform sense over a class of models worsens as the number of subsamples increases, leading to a version of the super-efficiency phenomenon. Then, we build a corrected pooled estimator that does not suffer from the super-efficiency phenomenon and allows for some heterogeneity in data. The new estimator essentially reverses the steps involved in constructing the above pooled estimator: we first smooth (by local averaging) on each subsample, and then isotonize the pooled smoothed data.

Title: ***Adaptive Confidence Bands in Multidimensional Shape-constrained Regression using Multiscale Methods***
Speaker: **Bodhisattva Sen**, Columbia University, United States of America
Co-author(s): Pratyay Datta
Schedule: *Wednesday 13 June, 09:30–10:00*
Abstract: We consider a multivariate continuous Gaussian white noise model with unknown regression function f, the parameter of interest. We propose a multivariate multiscale statistic and use it to construct confidence bands for f with guaranteed given coverage probability, assuming f is shape-constrained (multivariate isotonic or multivariate convex). The confidence bands are shown to be adaptive with respect to the smoothness of the underlying function, and to the intrinsic dimension (as opposed to the ambient dimension) of f. The confidence bands are also optimal in an appropriate sense, in some cases.
This is joint work with Pratyay Datta, a Ph.D, student at Columbia University.

Title: ***Interpreting the spectral embedding in finite mixture models***
Speaker: **Yuekai Sun**, University of Michigan, United States of America
Schedule: *Wednesday 13 June, 10:00–10:30*
Abstract: As datasets increase in heterogeneity, practitioners are turning to mixture models to account for the heterogeneity. Spectral clustering is a non-parametric approach that embeds the data into a low-dimensional space in a way that reveals the cluster structure in the data. We show that the spectral embedding in finite mixture models is, up to an orthogonal transformation, a perturbed version of the posterior probabilities of the unobserved labels. Based on this connection, we design a geometric algorithm to recover the latter from the former.

Invited Session IS-31 • Wednesday 13, 08:30-10:30 • Room Ravello

RECENT DEVELOPMENTS IN DEPENDENCE METRICS AND THEIR APPLICATIONS

Organizer(s): Shubhadeep Chakraborty (Texas A&M University)
Chair: Shubhadeep Chakraborty (Texas A&M University)

Title: ***Kernel Dependency Measures***
Speaker: **Bharath Sriperumbudur**, The Pennsylvania State University, United States of America

Co-author(s): Zoltan Szabo
Schedule: *Wednesday 13 June, 08:30–09:00*
Abstract: Given two random variables X and Y, it is a well-known fact that their independence can be characterized by the covariance between f(X) and g(Y) being zero for all bounded and continuous functions f and g. By choosing the functions f and g to belong to reproducing kernel Hilbert spaces (RKHS), the independence between X and Y can be characterized by the Hilbert-Schmidt norms of the cross-covariance operator induced by the reproducing kernels of these RKHSs. This is called the Hilbert-Schmidt Independence Criterion (HSIC) which measures the degree of dependence/independence between X and Y. In this work, we provide characterizations of kernels that guarantee: X and Y are independent if and only HSIC is zero. We also establish the connection of HSIC to distance covariance and investigate the applications of HSIC in independence testing.

Title: ***Kernel-based tests for joint independence***
Speaker: **Niklas Pfister**, ETH Zürich, Switzerland
Co-author(s): Peter Bühlmann, Bernhard Schölkopf, Jonas Peters
Schedule: *Wednesday 13 June, 09:00–09:30*
Abstract: In this talk we introduce a method for testing whether d possibly multivariate random variables, which may or may not be continuous, are jointly (or mutually) independent. Our method builds on ideas of the two-variable Hilbert–Schmidt independence criterion but allows for an arbitrary number of variables. We embed the joint distribution and the product of the marginals in a reproducing kernel Hilbert space and define the d-variable Hilbert–Schmidt independence criterion dHSIC as the squared distance between the 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.

Title: ***Regression with I-priors***
Speaker: **Wicher Bergsma**, London School of Economics, United Kingdom
Schedule: *Wednesday 13 June, 09:30–10:00*
Abstract: We propose an objective prior for a regression function which is assumed to lie in a reproducing kernel Hilbert space (RKHS).
Our prior, called I-prior, is defined as the distribution maximizing entropy subject to a suitable constraint based on the Fisher information for the regression function, and turns out to be Gaussian

with covariance kernel proportional to the Fisher information. The mean is chosen a priori (e.g., 0).
The I-prior has the intuitively appealing property that the more information is available about a linear functional of the regression function, the larger its prior variance, and, broadly speaking, the less influential the prior is on the posterior.

Title: ***Distance Metrics for Measuring Joint Dependence with Application to Causal Inference***

Speaker: **Shubhadeep Chakraborty**, Texas A&M University, United States of America

Co-author(s): Xianyang Zhang

Schedule: *Wednesday 13 June, 10:00–10:30*

Abstract: Many statistical applications require the quantification of joint dependence among more than two random vectors. In this work, we generalize the notion of distance covariance to quantify joint dependence among d >= 2 random vectors. We introduce the high order distance covariance to measure the so-called Lancaster interaction dependence. The joint distance covariance is then defined as a linear combination of pairwise distance covariances and their higher order counterparts which together completely characterize mutual independence. We further introduce some related concepts including the distance cumulant, distance characteristic function, and rank-based distance covariance. Empirical estimators are constructed based on certain Euclidean distances between sample elements. We study the large sample properties of the estimators and propose a bootstrap procedure to approximate their sampling distributions. The asymptotic validity of the bootstrap procedure is justified under both the null and alternative hypotheses. The new metrics are employed to perform model selection in causal inference, which is based on the joint independence testing of the residuals from the fitted structural equation models. The effectiveness of the method is illustrated via both simulated and real datasets.

Invited Session IS-32 • Wednesday 13, 08:30-10:30 • Room Furore

CHALLENGING PROBLEMS IN MODERN NONPARAMETRIC STATISTICS

Organizer(s): Lan Wang (University of Minnesota)
Chair: Lan Wang (University of Minnesota)

Title: ***A Composite Coefficient of Determination and Its Application in***

Ultrahigh Dimensional Variable Screening

Speaker: **Efang Kong**, Univeristy of Electronic Science and Technology, China, People's Republic of

Co-author(s): Yingcun Xia, Wei Zhong

Schedule: *Wednesday 13 June, 08:30–09:00*

Abstract: In this paper, we measure the dependence between two random variables using a composite coefficient of determination (CCD) of a set of nonparametric regressions. Each regression takes a consecutive binarization of one variable as response and the other variable as predictor. The measure is distribution-free and invariant to monotone variable transformations. It is thus robust to the data distributions and outliers. The composite coefficients of determination are comparable for different types of dependence. The CCD is estimated by kernel smoothing, and the estimator is root-n consistent. We utilize the CCD as a marginal utility to screen variables in ultrahigh dimensional data problems, and the sure screening properties are also justified. Extension of CCD to the mutual independence of multivariate random variables is also considered. Both comprehensive simulation studies and real data analysis demonstrate that the proposed measure is useful in testing the independence and in variable screening in regression.

Title: ***Goodness-of-fit tests in survival models with random effects***

Speaker: **Ingrid Van Keilegom**, KU Leuven

Co-author(s): Wenceslao Gonzalez Manteiga, Maria Dolores Martinez Miranda

Schedule: *Wednesday 13 June, 09:00–09:30*

Abstract: The paper deals with testing the functional form of the covariate effects in a Cox model with random effects, like for instance a shared frailty model. We assume that the responses are clustered and incomplete due to right censoring. The estimation of the model under the null (parametric covariate effect) and the alternative (non-parametric effect) is performed using full marginal likelihood. Under the alternative, the non-parametric covariate effects are represented using an orthogonal polynomial expansion. The test statistic is the likelihood ratio statistic, and its distribution is approximated using a bootstrap method. The proposed testing procedure is studied theoretically and through simulations.

Title: ***Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse***

Speaker: **Yanyuan Ma**, Penn State University, United States of America

Schedule: *Wednesday 13 June, 09:30–10:00*

Abstract: Tang, Little and Raghunathan (2003,Biometrika) considered a regression model with missing response, where the missingness

mechanism depends on the value of the response variable and hence is nonignorable. They proposed three different pseudolikelihood estimators, based on treatments to the probability distribution of the completely observed covariates. The first assumes the distribution of the covariate to be known, the second estimates this distribution parametrically, and the third estimates the distribution nonparametrically. While it is not hard to show that the second estimator is more efficient than the first, Tang et al. (2003) only conjectured that the third estimator is more efficient than the first two. In this paper, we investigate the asymptotic behavior of the third estimator by providing a closed-form representation of its asymptotic variance. We then prove that the third estimator is more efficient than the other two. Our result can be straightforwardly applied to missingness mechanisms that are more general than that in Tang et al (2003).

Title: ***Change point analysis in non-stationary processes - a mass excess approach***

Speaker: **Weichi Wu**, Ruhr University Bochum, Germany
Co-author(s): Dette Holger

Schedule: *Wednesday 13 June, 10:00–10:30*

Abstract: This paper considers the problem of testing if a sequence of means $(\mu_t)_{t=1,\ldots,n}$ of a non-stationary time series $(X_t)_{t=1,\ldots,n}$ is stable in the sense that the difference of the means μ_1 and μ_t between the initial time $t = 1$ and any other time is smaller than a given threshold, that is $|\mu_1 - \mu_t| \leq c$ for all \$t =1,...,n \$. A test for hypotheses of this type is developed using a bias-corrected monotone rearranged local estimator and asymptotic normality of the corresponding test statistic is established. As the asymptotic variance depends on the location of the critical roots of the equation $|\mu_1 - \mu_t| = c$ a new bootstrap procedure is proposed to obtain critical values and its consistency is established. As a consequence, we are able to quantitatively describe relevant deviations of a non-stationary sequence from its initial value. The results are illustrated by means of a simulation study and by analyzing data examples.

RECENT ADVANCES IN INFERENCE AND TESTING FOR STRUCTURAL CHANGE IN TIME SERIES MODELS

Organizer(s): Hira Koul (Michigan State University)
Chair: Liudas Giraitis (Queen Mary, University of London)

Title: ***Fitting a Two Phase Threshold Multiplicative Error Model***
Speaker: **Indeewara Perera**, Monash University, Australia
Co-author(s): Hira Koul
Schedule: *Wednesday 13 June, 08:30–09:00*
Abstract: The recent literature on time series analysis has devoted considerable attention to nonnegative time series, such as financial durations, realized volatility, and squared returns. The class of models, referred to as the multiplicative error models [MEM], is particularly suited to model such nonnegative time series. We develop a lack-of-fit test for fitting a two-phase threshold model for the conditional mean function in an MEM; in the case the jump size at the change-point is either fixed or tends to zero at a rate slower than the square root of the sample size. The test is asymptotically distribution free, consistent against a large class of fixed alternatives and has non-trivial asymptotic power against a class of nonparametric local alternatives. The proposed testing procedure can also be applied to a class of autoregressive conditional heteroscedastic threshold models. A test for the error distribution in a two phase threshold MEM is also proposed. We evaluate the test in a simulation study. The testing procedure is illustrated by using two data examples.

Title: ***Simultaneous multiple change-point and factor analysis for high-dimensional time series***
Speaker: **Matteo Barigozzi**, LSE, United Kingdom
Co-author(s): Haeran Cho, Piotr Fryzlewicz
Schedule: *Wednesday 13 June, 09:00–09:30*
Abstract: We propose the first comprehensive treatment of high-dimensional time series factor models with multiple change-points in their second-order structure. We operate under the most flexible definition of piecewise stationarity, and estimate the number and locations of change-points consistently as well as identifying whether they originate in the common or idiosyncratic components. Through the use of wavelets, we transform the problem of change-point detection in the second-order structure of a high-dimensional time series, into

the (relatively easier) problem of change-point detection in the means of high-dimensional panel data. Also, our methodology circumvents the di cult issue of the accurate estimation of the true number of factors in the presence of multiple change-points by adopting a screening procedure. We further show that consistent factor analysis is achieved over each segment de ned by the change-points estimated by the proposed methodology. In extensive simulation studies, we observe that factor analysis prior to change-point detection improves the detectability of change-points. Two empirical applications on financial and macroeconomic data are also presented.

Title: ***Testing for Structural Breaks via Ordinal Pattern Dependence***
Speaker: **Herold Dehling**, Ruhr-University Bochum, Germany
Co-author(s): Alexander Schnurr
Schedule: *Wednesday 13 June, 09:30–10:00*
Abstract: We propose new concepts to analyse and model the dependence structure between two time series. Our methods rely exclusively on the order structure of the data points. Hence, the methods are stable under monotone transformations of the time series and robust against small perturbations or measurement errors. We propose estimators for the ordinal dependence coefficient, and we calculate its asymptotic distribution. We derive a test for structural breaks within the dependence structure, and we investigate its asymptotic distribution.

Title: ***Inference on Time Series with Changing Mean and Variance***
Speaker: **Liudas Giraitis**, Queen Mary, University of London
Co-author(s): Violetta Dalla, Peter M Robinson
Schedule: *Wednesday 13 June, 10:00–10:30*
Abstract: The paper develops point estimation and large sample statistical inference with respect to a semiparametric model for time series with moving mean and unconditional heteroscedasticity. These two features are modelled nonparametrically, whereas autocorrelations are described by a short memory stationary parametric time series model. We first study the usual least squares estimate of the coefficient of the first-order autoregressive model based on constant but unknown mean and variance. We then consider standard Whittle estimates of a more general class of short memory parametric time series model, under otherwise more restrictive conditions. When the mean is correctly assumed to be constant, estimates that ignore the heteroscedasticity are again found to be consistent for the dependence parameters, and asymptotically normal with parametric rate, and inefficient. Allowing a slowly time-varying mean we resort to

trimming out of low frequencies to achieve the same outcome. Returning to finite order autoregression, nonparametric estimates of the varying mean and variance are given asymptotic justification, and forecasting formulae developed. Finite sample properties are studied by a small Monte Carlo simulations, and an empirical example is also included.

Invited Session IS-34 • Wednesday 13, 08:30-10:30 • Room Maiori

RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS

Organizer(s): Anirvan Chakraborty & Anil K. Ghosh (Ecole Polytechnique Federale de Lausanne (EPFL) & Indian Statistical Institute)

Chair: Todd Ogden (Columbia University)

Title: ***Nonparametric depth-based classification for functional data***

Speaker: **Germain Van Bever**, Université libre de Bruxelles, Brussels, Belgium

Co-author(s): Sami Helander, Stanislav Nagy, Lauri Viitasaari, Pauliina Ilmonen

Schedule: *Wednesday 13 June, 08:30–09:00*

Abstract: Functional observations are now pervasive in modern statistics and there exists several depth notions aiming at assessing the typicality of such data. Most of them, however, provide a center-outward ordering based either on location or, in rare instances, shape considerations. In this talk, we provide a new depth notion based on j-th order k-th moment integrated depths. The simplest instance, for $j = 1$ and $k = 1$, boils down to the mean of the univariate marginal halfspace depth distribution, a classical instance of integrated depth. For larger value of j, the proposed depth allows to consider shape and many other features of the functional data as measure of typicality. Moreover, the method can be applied to partially observed functions without extrapolation or interpolation. We also consider the classification that can be conducted using this new definition. Theoretical properties of the new approach are explored and several real data examples illustrate the concept and its classification performance.

Title: ***Spectral clustering of functional time series***

Speaker: **Anne M.N. van Delft**, Ruhr University Bochum, Germany

Co-author(s): Pramita Bagchi, Holger Dette

Schedule: *Wednesday 13 June, 09:00–09:30*

Abstract: Due to the surge of data storage techniques, the need for the

development of appropriate techniques to identify patterns and to extract knowledge from the resulting enormous data sets, which can be viewed as collections of dependent functional data, is of increasing interest in many scientific areas. We develop such a technique and introduce a spectral clustering algorithm for time series of functional data. First, we propose a measure to test equality of the spectral density operators of a collection of functional time series. The functional time series are neither supposed independent nor stationary. The measure is based on the aggregation of Hilbert-Schmidt differences of the individual time-varying spectral density operators. Under fairly general conditions, the asymptotic properties of the corresponding estimator are derived and asymptotic normality is established. The introduced statistic lends itself naturally to quantify (dis)-similarity between functional time series, which we subsequently exploit in order to build a spectral clustering algorithm. Our algorithm is the first of its kind in the analysis of nonstationary (functional) time series and enables to discover particular patterns by grouping together 'similar' series into clusters, thereby reducing the complexity of the analysis considerably. The algorithm is simple to implement and computationally feasible.

Title: ***Calculating a Generated Effect Modifier (GEM) for Treatment Selection based on Functional Data***

Speaker: **Todd Ogden**, Columbia University, United States of America
Co-author(s): Hyung Park, Eva Petkova, Thaddeus Tarpey

Schedule: *Wednesday 13 June, 09:30–10:00*

Abstract: A major goal in precision medicine is to make optimal patient-specific treatment decisions using data observed at baseline. For the treatment of neuropsychiatric disorders, available data may include clinical variables and measures of bevavioral/cognitive performance, as well as complex imaging data. We will present methods for (1) determining low-dimensional projections of all these data that are useful for describing differential treatment response; and (2) for estimating nonparametrically defined link functions based on these projections for each potential treatment. The resulting model can potentially provide powerful tools for precision medicine.

MODERN CHALLENGES IN MULTIVARIATE AND SPATIAL EXTREMES

Organizer(s): Raphael Huser (KAUST)
Chair: Raphael Huser (KAUST)

Title: ***Extremes of a Random Number of Multivariate Risks***
Speaker: **Simone Padoan**, Bocconi University, Italy
Co-author(s): Enkelejd Hashorva, Stefano Rizzelli
Schedule: *Wednesday 13 June, 08:30–09:00*
Abstract: Risk analysis in the area of insurance, financial and risk management is concerned with the study of the joint probability that multiple extreme events take place simultaneously. Extreme Value Theory provides tools for estimating such a type of probability. When aggregate data such as maxima of a random number of observations (risks), few results are available, and this is especially true in high dimensions. To fill this gap, we derive the asymptotic distribution of normalized maxima of multivariate risks, under appropriate conditions on the random number of observations. Of the latter we derive the extremal dependence which is for heavy-tailed scenarios stronger than the dependence among the usual multivariate maxima described in the classical theory. We establish the connection between the two dependence structures through random scaling and Pickands dependence functions. By means of the so-called inverse problem, we construct a semiparametric estimator for the extremal dependence of the unobservable data, starting from an estimator of the extremal dependence obtained with the aggregated data. We develop the asymptotic theory of the estimator and further explore by a simulation study its finite-sample performance.

Title: ***Inference for asymptotically independent samples of extremes***
Speaker: **Armelle Guillou**, Université de Strasbourg et CNRS, France
Co-author(s): Simone A. Padoan, Stefano Rizzelli
Schedule: *Wednesday 13 June, 09:00–09:30*
Abstract: An important topic of the multivariate extreme-value theory is to develop probabilistic models and statistical methods to describe and measure the strength of dependence among extreme observations. The theory is well established for data whose dependence structure is compatible with that of asymptotically dependent models. On the contrary, in many applications data do not comply with asymptotic dependent models and thus new tools are required. The authors bring

their contribution to this context, by considering a componentwise maxima approach and by proposing first a statistical test based on the classical Pickands dependence function to verify whether asymptotic dependence or independence holds. Then, they present a new Pickands dependence function to describe the extremal dependence under asymptotic independence. Finally, they estimate the latter and they study the main asymptotic properties of the estimator. Its performance is illustrated by a simulation study.

Title: ***Non-linear models for extremal dependence***
Speaker: **Valérie Chavez-Demoulin**, University of Lausanne, Switzerland
Schedule: *Wednesday 13 June, 09:30–10:00*
Abstract: In many applications, the extremal dependence structure of random vectors may vary with covariates. In this work, we develop a semi-parametric method for the estimation of non-stationary multivariate Pickands dependence functions. The proposed construction benefits from the flexible generalized additive modelling framework. The resulting estimator of the Pickands function is regularized, in the bivariate case, using constrained median smoothing B-splines and bootstrap confidence intervals are constructed. Finally, we present the results from a simulation study and apply the new methodology to a real dataset.
Joint work with Linda Mhalla and Philippe Naveau.

Title: ***Semi-parametric tail regression modeling framework for the estimation of extreme spatio-temporal quantiles***
Speaker: **Raphael Huser**, KAUST, Saudi Arabia
Co-author(s): Thomas Opitz, Haakon Bakka, Haavard Rue
Schedule: *Wednesday 13 June, 10:00–10:30*
Abstract: The goal of this work is to predict high daily precipitation quantiles for each month at observed and unobserved locations. To address this specific problem, we develop a general Bayesian generalized additive modeling framework tailored to estimate complex trends in marginal extremes observed over space and time. Our approach is based on a set of regression equations linked to the exceedance probability above a high threshold and to the size of the excess, the latter being modeled using the asymptotic generalized Pareto (GP) distribution suggested by Extreme-Value Theory. Latent random effects are modeled additively and semi-parametrically using Gaussian process priors, which provides high flexibility and interpretability. Fast and accurate estimation of posterior distributions is performed thanks to the Integrated Nested Laplace approximation (INLA), efficiently implemented in the R-INLA software. We show that the GP

distribution meets the theoretical requirements of INLA, and we then develop a penalized complexity prior specification for the tail index, which is a crucial parameter for extrapolating tail event probabilities. This prior concentrates mass close to a light exponential tail while still allowing heavier tails by penalizing the distance to the exponential distribution at a constant rate, thus avoiding overfitting.

Invited Session IS-36 • Wednesday 13, 08:30-10:30 • Room Procida

RECENT ADVANCES FOR HIGH-DIMENSIONAL, COMPLEX DATA

Organizer(s): Makoto Aoshima (University of Tsukuba)
Chair: Makoto Aoshima (University of Tsukuba)

Title: ***Consistency properties of regularized noise reduction methodology in high-dimensional settings***
Speaker: **Kazuyoshi Yata**, University of Tsukuba, Japan
Co-author(s): Makoto Aoshima
Schedule: *Wednesday 13 June, 08:30–09:00*
Abstract: In this talk, we consider principal component analysis (PCA) methods in high-dimensional settings. We first consider asymptotic properties of the conventional estimator for eigenvalues and PC directions. We show that the estimators are affected by the high-dimensional noise structure directly, so that they become inconsistent. In order to overcome such difficulties in a high-dimensional situation, Yata and Aoshima (2012, JMVA) developed a new PCA method called the noise-reduction (NR) methodology. We show that the NR method can enjoy consistency properties for both eigenvalues and PC directions in high-dimensional settings. The estimator of the PC directions by the NR method has a consistency property in terms of an inner product. However, it does not hold a consistency property in terms of the Euclid norm. With the help of a thresholding method, we modify the estimator and propose a regularized NR method. We show that it holds the consistency property of the Euclid norm.

Title: ***A Multivariate Regularized Estimator of High Dimensional Cross-Covariance Matrices***
Speaker: **Tommaso Proietti**, University of Rome Tor Vergata, Italy
Co-author(s): Alessandro Giovannelli
Schedule: *Wednesday 13 June, 09:00–09:30*
Abstract: The estimation of the (auto- and) cross-covariance matrices of respectively a stationary random process plays a central role in prediction

theory and time series analysis. When the dimension of the matrix is of the same order of magnitude as the number of observations and/or the number of time series, the sample crosss-covariance matrix provides an inconsistent estimator. In the univariate framework, we proposed an estimator based on regularizing the sample partial autocorrelation function, via a modified Durbin-Levinson algorithm that receives as an input the banded and tapered sample partial autocorrelations and returns a consistent and positive definite estimator of the autocovariance matrix. The talk presents and discusses the multivariate generalization, which is based on a regularized Whittle algorithm, shrinking the lag structure towards a finite order vector autoregressive system (by penalizing the partial canonical correlations), on the one hand, and shrinking the cross-sectional covariance towards a diagonal target, on the other. As the shrinkage intensity increases, the multivariate system converges to a set of unrelated univariate processes. We illustrate the merits of the proposal with respect to the problem of out of sample prediction and the estimation of the spectral density of high-dimensional time series.

Title: ***LIMITING LAWS FOR DIVERGENT SPIKED EIGENVALUES***
Speaker: **Guangming Pan**, Nanyang Technological University, Singapore
Schedule: *Wednesday 13 June, 09:30–10:00*
Abstract: We study the asymptotic distributions of the spiked eigenvalues and the largest nonspiked eigenvalue of the sample covariance matrix under a general covariance matrix model with divergent spiked eigenvalues, while the other eigenvalues are bounded but otherwise arbitrary. The limiting normal distribution for the spiked sample eigenvalues is established. It has distinct features that the asymptotic mean relies on not only the population spikes but also the nonspikes and that the asymptotic variance in general depends on the population eigenvectors. In addition, the limiting Tracy-Widom law for the largest nonspiked sample eigenvalue is obtained. Estimation of the number of spikes and the convergence of the leading eigenvectors are also considered. The results hold even when the number of the spikes diverges.

Title: ***High-dimensional statistical analysis under spiked models***
Speaker: **Makoto Aoshima**, University of Tsukuba, Japan
Schedule: *Wednesday 13 June, 10:00–10:30*
Abstract: Any high-dimensional data is classified into two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. In actual high-dimensional data, one often finds a non-sparse and low-rank structure which contains strongly spiked eigenvalues.

That structure fits the SSE model. It can be noted that, under the SSE model, asymptotic normality of high-dimensional statistics is not valid because it is heavily influenced by strongly spiked eigenvalues. In order to give a unified treatment of both the SSE models and non-SSE models, data transformation techniques that transform the SSE models to the non-SSE models were developed. In this way, strongly spiked eigenvalues are accurately detected by using new PCA-type techniques. With the transformed data, one can create a new statistic which can ensure high accuracy for inferences by using asymptotic normality even under the SSE models. In this talk, the new techniques to handle high-dimensional data will be demonstrated in two-sample problems and classification problems.

Contributed Session CS-16
• Wednesday 13, 11:00-12:00 • Room Ravello

REGRESSION MODELS

Chair: Sabyasachi Chatterjee (University of Illinois at Urbana-Champaign)

Title: ***Estimation of Jump Discontinuities in Regression: a Generalized Reflection Approach***

Speaker: **Sihong Xie**, University of Colorado, United States of America

Schedule: *Wednesday 13 June, 11:00–11:20*

Abstract: We propose a new estimator for the size of a jump discontinuity on a nonparametric regression. The basic idea of our estimator is to extend the regressions on both sides of the discontinuity using the extension of Hestenes [1941]. These two extended regressions are then estimated and used to estimate the jump discontinuity. The inspiration for our method comes from recent work by Mynbaev and Martins-Filho [2017], where a simple and elegant solution to boundary problems in density estimation is obtained using the same extension principle. Our work provides a class of jump estimators that are easy to construct using classical kernels and constant bandwidth over the entire domain of the regression. Focusing on properties at boundary points, we provide bias, variance and asymptotic distribution of our estimators and compare them with those of Nadaraya-Watson (NW) and local linear (LL) estimators. We conduct extensive Monte Carlo simulations to contrast the finite sample performance of our estimators with that of the NW and LL estimators and investigate

the sources of bias at the discontinuous point. Finally, we apply our estimators to a study by Litschig and Morrison [2013] to illustrate empirical applicability of our estimators in regression discontinuity designs.

Title: ***"Is completeness necessary? Estimation in non-identified linear models"***

Speaker: **Andrii Babii**, University of North Carolina at Chapel Hill, United States of America

Co-author(s): Jean-Pierre Florens

Schedule: *Wednesday 13 June, 11:20–11:40*

Abstract: This paper studies non-identified ill-posed inverse models with estimated operator. Leading examples are the nonparametric IV regression and the functional linear IV regression. We argue that identification of infinite-dimensional parameters is less crucial than identification of finite-dimensional parameters. We show that in the case of identification failures, a very general family of continuously-regularized estimators is consistent for the best approximation of the parameter of interest and obtain $L_2$and L_∞ finite-sample risk bounds. This class includes Tikhonov, iterated Tikhonov, spectral cut-off, and Landweber-Fridman as special cases. We show that in many cases the best approximation coincides with the structural parameter and can be a useful and tractable object to infer relation between economic variables otherwise. Unlike in the identified case, estimation of the operator may have a non-negligible impact on the estimation accuracy and inference. We develop inferential methods for linear functionals in non-identified models as well as honest uniform confidence sets for the best approximation. Lastly, we demonstrate the discontinuity in the asymptotic distribution for extreme cases of identification failures where we observe a degenerate U-statistic s asymptotics.

Title: ***Isotonic Regression in General Dimensions***

Speaker: **Sabyasachi Chatterjee**, University of Illinois at Urbana-Champaign, United States of America

Co-author(s): Roy Han, Tengyao Wang, Richard Samworth

Schedule: *Wednesday 13 June, 11:40–12:00*

Abstract: We study the least squares regression function estimator over the class of real-valued functions on $[0,1]^d$ that are increasing in each coordinate. For uniformly bounded signals and with a fixed, cubic lattice design, we establish that the estimator achieves the minimax rate of order $n - \min\{2/(d+2), 1/d\}$ in the empirical L_2 loss, up to poly- logarithmic factors. Further, we prove a sharp oracle inequality, which reveals in particular that when the true regression function is

piece- wise constant on k hyper rectangles, the least squares estimator enjoys a faster, adaptive rate of convergence of $(k/n)^{min(1,2/d)}$, again up to poly-logarithmic factors. Previous results are confined to the case $d \leq 2$. Finally, we establish corresponding bounds (which are new even in the case $d = 2$) in the more challenging random design setting. There are two surprising features of these results: first, they demonstrate that it is possible for a global empirical risk minimisation procedure to be rate optimal up to poly-logarithmic factors even when the corresponding entropy integral for the function class diverges rapidly; second, they indicate that the adaptation rate for shape-constrained estimators can be strictly worse than the parametric rate.

Contributed Session CS-17
• Wednesday 13, 11:00-12:00 • Room Furore

SMOOTHING TECHNIQUES I

Chair: Juan-Carlos Pardo-Fernandez (Universidade de Vigo)

Title: ***Asymptotic behavior of a local-polynomial estimator of copula density function under random censoring***
Speaker: **Taoufik Bouezmarni**, Université de sherbrooke, Canada
Co-author(s): Yassir Rabhi
Schedule: *Wednesday 13 June, 11:00–11:20*
Abstract: This paper addresses the nonparametric estimation of the copula density and measures of dependence for right-censored data. We propose a local-polynomial estimator for the copula density that overcomes boundary bias and has a stable variance near 0 and 1. To study its asymptotic behavior theory, we first investigate the oscillation behavior of a bivariate distribution estimator, for right-censored data, and derive its i.i.d. representation with a faster remainder-term's rate. Then, we obtain an i.i.d representation for a copula cdf estimator and establish a functional CLT result on the copula density estimator. In addition, we introduce estimators for the Kendall and Spearman measures of dependence and establish their weak convergence. The proposed method is then applied to analyze a Time-to-Event data on survival with heart-transplant.

Title: ***Density estimation via nonparametric penalized likelihood***
Speaker: **Federico Ferraccioli**, Università degli Studi di Padova, Italy
Co-author(s): Laura M. Sangalli, Livio Finos
Schedule: *Wednesday 13 June, 11:20–11:40*

Abstract: A nonparametric likelihood approach to multivariate density estimation is considered. A regularization term based on the Laplace operator of the logarithm of the density controls the smoothness of the estimate and prevents degenerate solutions. The Laplacian is a measure of local curvature that is invariant with respect to Euclidean transformations of spatial coordinates, and therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. The complexity of the estimation problem is tackled by means of a finite element formulation, that allows great flexibility and computational tractability. The model is suitable for any type of bounded planar domain and can be generalized to non-Euclidean settings. The flexibility of the nonparametric likelihood approach allows the estimation of very complex structures, a fundamental feature in research areas such as density based clustering. Based on the proposed method, a clustering procedure stimulated by Morse theory is discussed. Within this framework, a permutation-based procedure for one and two samples hypothesis testing is also introduced.

Title: ***Asymptotic distribution-free tests for semiparametric regressions***
Speaker: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-author(s): Juan-Carlos Escanciano, Ingrid Van Keilegom
Schedule: *Wednesday 13 June, 11:40–12:00*
Abstract: In this talk, a new general methodology for constructing nonparametric and semiparametric asymptotically distribution-free (ADF) tests for semiparametric hypotheses in regression models will be presented. Classical tests based on the difference between the estimated distributions of the restricted and unrestricted regression errors are not ADF. Here, we introduce a novel transformation of this difference that leads to ADF tests with well-known critical values. The proposed methodology can be applied to independent observations or dependent data coming from a strictly stationary process. The general methodology is illustrated with applications to testing for parametric models against nonparametric or semiparametric alternatives, and semiparametric constrained mean-variance models. Several Monte Carlo studies show that the finite sample performance of the proposed tests is satisfactory in moderate sample sizes.

Contributed Session CS-18
• Wednesday 13, 11:00-12:00 • Room Positano

CENSORED DATA

Chair: Elodie BRUNEL (University of Montpellier)

Title: ***A new lack-of-fit test for quantile regression with censored data***
Speaker: **Mercedes Conde-Amboage**, Universidade de Santiago de Compostela, Spain
Co-author(s): Ingrid Van Keilegom, Wenceslao González Manteiga
Schedule: *Wednesday 13 June, 11:00–11:20*
Abstract: Although mean regression is still a traditional benchmark in regression studies, the quantile approach is receiving increasing attention, because it allows a more complete description of the conditional distribution of the response given the covariate, and it is more robust to deviations from error normality. That is, while classical regression gives only information on the conditional expectation, quantile regression extends the viewpoint on the whole conditional distribution of the response variable. For all that, quantile regression is a very useful statistical technology for a large diversity of disciplines. A new lack-of-fit test for quantile regression models will be presented, where the response variable is right-censored. The test is based on the cumulative sum of residuals, that is, extends the ideas of He and Zhu (2003) 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 it is consistent. To approximate the critical values of the test, a bootstrap mechanism is used. A simulation study was carried out that show the good properties of the new test versus other tests available in the literature.

Title: ***General regression model for the subdistribution of a competing risk under left-truncation and right-censoring***
Speaker: **Anna Bellach**, Fred Hutch Cancer Research Center, US
Co-author(s): Michael R Kosorok, Peter Gilbert, Jason P Fine
Schedule: *Wednesday 13 June, 11:20–11:40*
Abstract: Left-truncation poses additional challenges for the analysis of complex time to event data. We propose a general semiparametric regression model for left-truncated and right-censored competing risks data. Targeting the subdistribution hazard, our parameter estimates are directly interpretable with regard to the cumulative

incidence function. Our approach accommodates external time dependent covariate effects on the subdistribution hazard. We establish consistency and asymptotic normality of the estimators and propose a sandwich estimator of the variance. In comprehensive simulation studies we demonstrate a solid performance of the proposed method, thereby comparing the sandwich estimator to the inverse Fisher information. Applying the new method to HIV-1 vaccine efficacy trial data we investigate how participant factors associate with the time from adulthood until HIV-1 infection.

Title: ***Nonparametric survival function estimation for interval censoring case 2***

Speaker: **Elodie Brunel**, University of Montpellier, France

Co-author(s): Olivier Bouaziz, Fabienne Comte

Schedule: *Wednesday 13 June, 11:40–12:00*

Abstract: Our aim is to propose a new strategy of estimation for the survival function, associated to a survival time subject to interval censoring case 2. Our method is based on a least-squares contrast of regression type with parameters corresponding to the coefficients of the development of the survival function on an orthonormal basis. We obtain a collection of projection estimators where the dimension of the projection space has to be adequately chosen via a model selection procedure. For compactly supported bases, we obtain adaptive results leading to general nonparametric rates. However, our results can be used for non compactly supported bases, a true novelty in regression setting, and we use specifically the Laguerre basis which is supported on the positive real line and thus well suited when nonnegative random variables are involved in the model. Simulation results comparing our proposal with previous strategies show that it works well in a very general context. A real data set is considered to illustrate the methodology.

Contributed Session CS-19
• Wednesday 13, 11:00-12:00 • Room Procida

INFERENCE FOR LARGE SCALE PROBLEMS

Chair: Vladimir Pastukhov (Lund University)

Title: ***Change-point detection in a dynamic panel data model***

Speaker: **Zuzana Praskova**, Charles University, Faculty of Mathematics and Physics, Czech Republic

Schedule: *Wednesday 13 June, 11:00–11:20*

Abstract: A panel data model with lagged dependent variables and unobserved individual effects is considered and a procedure to detect change in coefficients of lagged variables is proposed. Test statistic to detect change is based on quasi-likelihood scores and quasi-maximum likelihood estimators. Asymptotic properties of the test statistic are studied in case that both the number of panels and the number of observations is sufficiently large. Asymptotic distribution of the test statistic under the null hypothesis is a functional of a Gaussian process. Various bootstrap variants to the asymptotic distribution that can approximate critical values is proposed and their efficiency is discussed.

Title: ***Collective Matrix Completion***

Speaker: **Mokhtar Z. Alaya**, Modal'X Laboratory, Paris Nanterre University, France

Co-author(s): Olga Klopp

Schedule: *Wednesday 13 June, 11:20–11:40*

Abstract: Matrix completion aims to reconstruct a high-dimensional data matrix based on a small number of entries possibly observed with noise. Normally, matrix completion problems consider a single matrix, which can be, for example, a rating matrix in recommendation systems. However in common practical situations, data is often obtained from a collection of matrices that covered multiple sources (domains) rather a single one. In this work, we consider the problem of collective matrix completion, where multiple and heterogeneous matrices, such as count, binary, continuous, etc., are reconstructed. We first investigate the setting where, for each source, the matrix entries are sampled from exponential family distributions with a matrix parameter to be estimated. In a second setting, we deal with the prediction problem of the collective matrix using loss functions appropriated to the heterogeneous data-types. The estimation procedures are based on the penalized nuclear norm procedure of the whole collective matrix. We prove that the proposed estimators achieve fast rates of convergence under the two considered settings.

Title: ***The asymptotic distribution of the isotonic regression estimator over a countable pre-ordered set***

Speaker: **Vladimir Pastukhov**, Lund University, Sweden

Co-author(s): Dragi Anevski

Schedule: *Wednesday 13 June, 11:40–12:00*

Abstract: We study the isotonic regression estimator over a general countable pre-ordered set. We obtain the limiting distribution of the estimator

and study its properties: It is proved that, under some general assumptions, the limiting distribution of the isotonised estimator is given by the concatenation of the separate isotonic regressions of the restrictions of an underlying estimator's asymptotic distribution to the comparable level sets of the underlying estimator's probability limit. Also, we show that the isotonisation preserves the rate of convergence of the underlying estimator. We apply these results to the problems of estimation of a bimonotone regression function and estimation of a bimonotone probability mass function.

Special Invited Session SIS-05 • Wednesday 13, 11:00-12:00 • Auditorium Tafuri

INTERCEPT ESTIMATION IN (NON)-ADDITIVE SEMIPARAMETRIC SELECTION MODELS

by Valentina Corradi[1] **& Wiji Arulampalam**[2] **& Daniel Gutknecht**[3]
[1]University of Surrey, UK; [2]Warwick University, UK; [3]Mannheim University, Germany

Speaker: Valentina Corradi
Chair: Walter Distaso (Imperial College London)

Abstract: This paper develops new estimators of the intercept for semiparametric sample selection models, which allows to recover treatment effects with these models in non-experimental settings. In the linear additive case, we introduce a local polynomial estimator which achieves the 'optimal' univariate nonparametric rate and improves over existing estimators as it does not require independence of the selection error and the regressor(s), and may be implemented using a data-driven bandwidth procedure. In the additive nonlinear and in the multiplicative model, estimation of the intercept has, to the best of our knowledge, not yet been addressed. We aim at filling this gap, deriving a bias corrected nonlinear least squares estimator. Depending on the properties of the (conditional) distribution of the index-valued instrument vector, this estimator may converge at a univariate nonparametric rate as well. Re-visiting a famous training data example, we find that applying our method to the non-experimental control group recovers the experimental average treatment effect.

Invited Session IS-37 • Wednesday 13, 12:00-14:00 • Room Vietri

GRAPH BASED STATISTICAL METHODS FOR HIGH DIMENSIONAL DATA

Organizer(s): Anil K. Ghosh & Subhadeep Mukhopadhyay (Indian Statistical Institute & Temple University)

Chair: Subhadeep Mukhopadhyay (Temple University)

Title: ***Performance analysis and robustness for dimension reduction***

Speaker: **Shojaeddin Chenouri**, University of Waterloo, Canada

Co-author(s): Jiaxi Liang, Christopher Small

Schedule: *Wednesday 13 June, 12:00–12:30*

Abstract: Information in the data often has far fewer degrees of freedom than the number of variables encoding the data. Dimensionality reduction attempts to reduce the number of variables used to describe the data. There are several dimensionality reduction methods available in the literature for linear and nonlinear manifolds. Each method works only under certain underlying assumptions. There is no universal agreement on how to assess and compare the performance of these different methods, and their robustness properties have not been studied. We attempt to discuss these issues and provide some answers. We introduce a goodness measure called local Spearman correlation for assessing performance and then employ it to define types of influence function and breakdown point to study the robustness of dimensionality reduction methods.

Title: ***New two-sample tests based on adjacency***

Speaker: **Hao Chen**, University of California, Davis, USA

Schedule: *Wednesday 13 June, 12:30–13:00*

Abstract: Two-sample tests for multivariate data and non-Euclidean data are widely used in many fields. We study a nonparametric testing procedure that utilizes graphs representing the similarity among observations. It can be applied to any data types as long as an informative similarity measure on the sample space can be defined. Existing tests based on a similarity graph lack power either for location or for scale alternatives. A new test is proposed that utilizes a common pattern overlooked previously, and it works for both types of alternatives. The test exhibits substantial power gains in simulation studies. Its asymptotic permutation null distribution is derived and shown to work well under finite samples, facilitating its application to large data sets. Another new test statistic will also be discussed that addresses the problem of the classic test of the type under unequal

sample sizes. Both tests are illustrated on an application of comparing networks under different conditions.

Title: ***Modifications of Some Graph-based Two-sample Tests for High-dimensional Data***

Speaker: **Soham Sarkar**, Indian Statistical Institute, Kolkata, India

Co-author(s): Rahul Biswas, Anil K. Ghosh

Schedule: *Wednesday 13 June, 13:00–13:30*

Abstract: We consider the problem of testing for the equality of two d-dimensional distributions F and G based on two independent samples $x_1, \ldots, x_m \sim F$ and $y_1, \ldots, y_n \sim G$, where d is much larger than $m+n$. Most of the classical tests (e.g., Hotelling's T^2 test) cannot be used in such high dimension, low sample size (HDLSS) setup. Some popular graph-based nonparametric tests, e.g., the multivariate run test or the nearest neighbor test, can be used in such situations. But these tests suffer in the HDLSS setup due to the *concentration of the Euclidean distance*. We propose simple modifications of some of these existing tests using some new dissimilarity indices, which can be adequately used in the HDLSS context. These modified tests outperform many popular nonparametric tests in a wide variety of problems. We establish the high dimensional consistency of these modified tests under appropriate regularity conditions and demonstrate their usefulness through the analysis of several simulated and real datasets.

Title: ***Network Homogeneity via Spectral Unimodality: Concept, Theory and Application***

Speaker: **Sandipan Roy**, University College London, United Kingdom

Co-author(s): Subhadeep Mukhopadhyay

Schedule: *Wednesday 13 June, 13:30–14:00*

Abstract: Detecting deviation from the baseline model is a fundamental statistical exercise and is often the first step for assessing possible hidden *structure*. This paper introduces a new hypothesis testing framework to test whether a given graph is *homogeneous*. In a real world-network it is crucial to understand the network homogeneity as it reveals the patterns of relationships among the individual nodes (actors). Two particularly noteworthy aspects of our idea are :(1) reformulating graph homogeneity test of graph spectra, which brings unexpected theoretical simplicity; (2) developing an exploratory nonparametric algorithm and diagnostic tools for easy interpretation. Finally, we apply this idea for change-point detection in graphs and two sample high-dimensional testing. We validate our proposed approach with several synthetic experiments. We also apply our proposed approach to detect number of communities in different real

data applications and show superior performance of our approach compared to existing methods.

Invited Session IS-38 • Wednesday 13, 12:00-14:00 • Room Ravello

NONPARAMETRIC AND SEMIPARAMETRIC TEST AND MODEL ADEQUECY IN HIGH DIMENSIONAL DATA

Organizer(s): Haiyan Wang & Jiguo Cao (Kansas State University & Simon Fraser University)

Chair: Haiyan Wang (Kansas State University)

Title: ***A high dimensional nonparametric test to compare time series***

Speaker: **Lei Jin**, Texas A&M University Corpus Christi, United States of America

Schedule: *Wednesday 13 June, 12:00–12:30*

Abstract: In this paper, a high dimensional non-parametric test is proposed to compare stationary time series, in terms of the second order dynamics. To develop test, a high dimensional regression model is used to model the difference between two spectral densities. A test statistic is proposed based on lasso fitted values. The asymptotic null distribution and the consistency of test are obtained. An extensive simulation study illustrates the validity of the asymptotic result and finite sample properties. The proposed test is found to perform well under different situations. An application to damage detection for a mechanical system using vibration data is discussed.

Title: ***Consistency of the estimator in Spatial Regression with PDE penalization***

Speaker: **Eleonora Arnone**, Politecnico di Milano, Italy

Co-author(s): Alois Kneip, Fabio Nobile, Laura Maria Sangalli

Schedule: *Wednesday 13 June, 12:30–13:00*

Abstract: We study the consistency of the estimator in Spatial Regression with Partial Differential Equation penalization method (SR-PDE). SR-PDE is a technique for the estimation of a spatial dependent field over a two-dimensional complex domain from pointwise noisy observations when prior information on the field is available in form of a PDE. The consistency is studied both for the estimator in the infinite dimensional setting and for the discrete estimator obtained with finite elements method. Bias and variance of the estimator are analyzed with respect to the sample size and the value of the smoothing parameter. It is shown that optimal rates of convergence

can be reached for the mean squared error in the L^2 and discrete norm when the number of observations goes to infinity. Simulation studies to verify the convergence rates are performed in a simple setting.

Title: ***A two-sample test of equality of means in high dimensional data***
Speaker: **Haiyan Wang**, Kansas State University, United States of America
Co-author(s): Huaiyu Zhang
Schedule: *Wednesday 13 June, 13:00–13:30*
Abstract: This research is interested in testing equality of two sample means in high dimensional data in which the sample sizes may be much less than the dimension. Improvement still can be achieved despite significant effort in recent literature that modifies the Hotelling's T2-statistics by either bypassing the estimation of high dimensional covariance matrices (cf. Chen & Qin 2010 Annals of Stat., Srivastava et al. 2013 JMVA, Gregory et al. 2015 JASA) or estimating the precision matrix after imposing sparseness condition (cf. Cai et al. 2014 JRSSB). Here we present a new test statistic that is particularly powerful when the correlation between components of the data vector reduces as the separation of the component indices increases. The limiting distribution of the test statistic and power of the test are studied. Simulation results will be presented to show the numerical performance of the test and to compare with other tests in the literature.

Invited Session IS-39 • Wednesday 13, 12:00-14:00 • Room Furore

SEMIPARAMETRIC INFERENCE

Organizer(s): Ursula U. Müller & Valentin Patilea (Texas A&M University & ENSAI)
Chair: Ursula U. Müller (Texas A&M University)

Title: ***Illumination problems in digital images. A statistical point of view***
Speaker: **Myriam Anne Vimond**, CREST (Ensai), France
Co-author(s): Ségolen Geffray, Nicolas Klutchnikoff
Schedule: *Wednesday 13 June, 12:00–12:30*
Abstract: We focus on the statistical treatment of illumination artefacts on digital images in the presence of an additional random noise. We assume that this artefact consists of "smooth" variations of the intensity of the signal of interest R. Such an assumption is classically modelled using a function L which acts in a multiplicative way on R. Our goal is to estimate R from observations of a random

variable Y which obeys the regression model $Y = RL + \varepsilon$. Our main contribution lies in the derivation of a new estimator of R which is shown to be consistent under suitable identifiability and regularity conditions. The accuracy of this new estimation procedure is studied from a theoretical point of view through the rate of convergence of the uniform risk. Applications to real Scanning Electron Microscopy images are presented, as well as a qualitative study of the performances of our method with respect to other image processing techniques.

Title: ***The nonparametric location-scale mixture cure model***
Speaker: **Justin Andrew Chown**, Ruhr-Universitat Bochum, Germany
Co-author(s): Cedric Heuchenne, Ingrid Van Keilegom
Schedule: *Wednesday 13 June, 12:30–13:00*
Abstract: We propose completely nonparametric methodology to investigate location–scale modelling of two–component mixture cure models. These models are interesting because the responses are only indirectly observable due to the presence of censoring and the presence of so–called long–term survivors that are always censored. We investigate an estimator of the error distribution function that has not been considered before in the literature. This estimator depends on a bandwidth. When this bandwidth belongs to a certain range of undersmoothing bandwidths, the proposed estimator does not appear sensitive to the choice of bandwidth. This suggests that a computationally costly bandwidth selection procedure is unnecessary to obtain an effective estimator of the error distribution, and that a simpler rule-of-thumb approach can be used instead. A simulation study investigates the finite sample properties of our approach, and the methodology is illustrated using data obtained to study the behavior of distant metastasis in lymph-node-negative breast cancer patients.

Title: ***Efficient semiparametric estimation in time-varying regression models***
Speaker: **Lionel Truquet**, ENSAI, France
Schedule: *Wednesday 13 June, 13:00–13:30*
Abstract: We study semiparametric inference in some linear regression models with time-varying coefficients, dependent regressors and dependent errors. Using the locally stationary framework and the functional dependence measure of Wu (2005), we use a profile least-squares approach for estimating non time-varying parameters at the root-n rate. For a time-varying AR process with exogenous covariates and conditionally Gaussian errors, we derive a notion of efficient

information matrix from a convolution theorem adapted to triangular arrays. For independent but non identically distributed Gaussian errors, we construct an asymptotically efficient estimator in a semiparametric sense. Our results fill some gaps for estimating such semi-varying regression models.

Title: ***Semi-parametric transformation boundary regression models***
Speaker: **Charles Tillier**, University of Hamburg, Germany
Co-author(s): Natalie Neumeyer, Leonie Selk
Schedule: *Wednesday 13 June, 13:30–14:00*
Abstract: In mean and quantile regression models transformations of the response have been applied in order to obtain simpler models or gain efficiency in the statistical inference. Parametric classes such as Box-Cox power transformations and generalizations of those are common in practise. We will consider such parametric transformations of the response in the context of non-parametric regression models with one-sided errors (boundary regression) in order to attain independence of the error distribution from the covariates. The estimator of the transformation is based on a minimum distance approach. Consistency is shown and the small sample behaviour is demonstrated in a simulation study.

Invited Session IS-40 • Wednesday 13, 12:00-14:00 • Room Positano

ESTIMATION METHODS IN TIME SERIES: FREQUENCY DOMAIN AND MULTIVARIATE ANALYSIS

Organizer(s): Alessandra Luati (University of Bologna)
Chair: Alessandra Luati (University of Bologna)

Title: ***Likelihood Based Inference in Dynamic Mixture Cointegrated VAR Models***
Speaker: **Paolo Paruolo**, Joint Research Centre, European Commission, Italy
Schedule: *Wednesday 13 June, 12:00–12:30*
Abstract: We consider likelihood-based asymptotic inference in a general class of regime-switching, or dynamic mixture, cointegrated vector error correction models. This framework allows for epochs of non-stationary behavior, asymmetric error correction and switching error covariance; this extends previously-introduced classes of processes. Unlike previous results on non-linear switching cointegrated models, we discuss asymptotic inference on all parameters, including the cointegrating vectors and switching covariances. To do so, we

introduce a new functional central limit theory for non-stationary switching processes, find explicit conditions for existence of moments and derive limiting distributions. The small sample properties are investigated through a simulations study.
(joint paper with Emil Nejstgaard and Anders Rahbek)

Title: ***Orthogonal samples for estimators in time series and spatial data***
Speaker: **Suhasini Subba Rao**, Texas A&M University, United States of America
Schedule: *Wednesday 13 June, 12:30–13:00*
Abstract: Inference for statistics of a stationary time series or spatial data often involve nuisance parameters and sampling distributions that are difficult to estimate. In this talk, we propose the method of orthogonal samples, which can be used to address some of these issues. For a broad class of statistics, an orthogonal sample is constructed through a slight modification of the original statistic, such that it shares similar distributional properties as the centralised statistic of interest. We use the orthogonal sample to estimate nuisance parameters of weighted average periodogram estimators and L_2-type spectral statistics. Further, the orthogonal sample is utilized to estimate the finite sampling distribution of various test statistics under the null hypothesis. The proposed method is simple and computationally fast to implement. The viability of the method is illustrated with various simulations.

Title: ***Generalised autocovariances and spectral estimators***
Speaker: **Alessandra Luati**, University of Bologna, Italy
Co-author(s): Tommaso Proietti, Stefano Grassi
Schedule: *Wednesday 13 June, 13:00–13:30*
Abstract: The basic theory of frequency domain estimation of generalised autocovariances and related quantities for stationary processes is reviewed.The attention is then focused on inference for spectral estimators in locally stationary processes.
A class of models for the time-varying spectra of locally stationary processes is introduced. The class depends on a power parameter that applies to the spectrum so that it can be locally represented by a finite Fourier polynomial. The coefficients of the polynomial are time varying generalised cepstral coefficients. Estimation is based on the pre-periodogram and Whittle likelihood.

ADVANCES IN OBJECT DATA ANALYSIS I

Organizer(s): Vic Patrangenaru (Florida State University)
Chair: Leif Ellingson (Texas Tech University)

Title: ***Representing uncertainty in Space Situational Awareness***
Speaker: **John T Kent**, University of Leeds, United Kingdom
Co-author(s): Shambo Bhattacharjee
Schedule: *Wednesday 13 June, 12:00–12:30*
Abstract: There are many thousands of "resident space objects" (RSOs) in orbit about the earth, including both operational satellites and large and small pieces of debris. Tracking these objects is a main theme in Space Situational Awareness, and a key objective is ensure that operational satellites do not collide with other RSOs. For statistical purposes it is convenient to take the "data object" to be information in the state vector of the RSO, i.e. the 6-dimensional vector of position and velocity. An important step in the tracking problem is to propagate an initial uncertainty in the state to a later time. However, even if the initial uncertainty is Gaussian, the propagated uncertainty can quickly become non-Gaussian in Euclidean coordinates. The reason is that the uncertainty in the propagated state becomes spread out along the (curved) orbital path. A new "adapted structural" (AST) coordinate system has been developed to deal with this issue. In AST coordinates the propagated state distribution remains approximately Gaussian under a wide range of conditions. Using AST coordinates, a Kalman-like filter has been developed to iteratively update the state distribution from a sequence of optical measurements.

Title: ***Smeary Limit Theorems***
Speaker: **Stephan Huckemann**, Universtät Göttingen, Germany
Co-author(s): Benjamin Eltzner
Schedule: *Wednesday 13 June, 12:30–13:00*
Abstract: We recall Frechet means for manifolds, the central limit theorem from the seminal paper by Bhattacharya and Patrangenaru (2005) and extend this result to more general scenarios involving "smeary" asymptotics. It turns out that smeary limit theorems may have unprecedented implications on finite sample asymptotics also in the classical nonsmeary case. This effect can increase with dimension.

Title: ***A Covariance-based Representation of Protein Binding Sites for Model-based Classification***

Speaker: **Leif Ellingson**, Texas Tech University, United States of America
Co-author(s): G. A. I. C. Premarathna
Schedule: *Wednesday 13 June, 13:00–13:30*
Abstract: It is known that a protein's biological function is in some way related to its physical structure. However, it remains an open challenge to predict a protein's function from its structure. Many approaches to this problem involve computing univariate descriptors for a protein's structure, resulting in much information being lost. Another option is to utilize higher level object representations of this information. In both cases, a problem that must be overcome is that the molecular structures of proteins are provided with respect to arbitrary coordinate systems, which commonly results in the need to perform computationally expensive pairwise alignments of the protein structures. The pairwise comparisons obtained from such procedures are often also difficult to use for model-based classification. The methodology considered here represents protein binding site structures as covariance matrices in a manner that is invariant to rotation and translation, thus eliminating the need for pairwise alignment. It also simultaneously reduces information loss compared to univariate representations and allows for model-based classification using Mahalanobis distance. This approach is illustrated for two benchmark data sets. If time permits, the impact of the choice of distance on the space of symmetric positive definite matrices will be discussed.

Invited Session IS-42 • Wednesday 13, 12:00-14:00 • Room Procida

RECENT ADVANCES IN NONPARAMETRIC SURVIVAL ANALYSIS
Organizer(s): Arnold Janssen (Heinrich Heine University)
Chair: Arnold Janssen (Heinrich Heine University)

Title: ***A permutation test for proportional hazards under censored data***
Speaker: **Marc Ditzhaus**, Heinrich Heine University Duesseldorf, Germany
Co-author(s): Arnold Janssen
Schedule: *Wednesday 13 June, 12:00–12:30*
Abstract: Let us consider a two-sample survival setting under right censoring. For this and more general settings the Cox model is very popular and often supposed to be true in practice. Hence, tests to check its model assumptions (proportional hazard rates) are needed. We suggest a studentized permutation test for this testing problem. Using discrete

martingale techniques we prove convergence of the test statistic under the null. Beside the asymptotic theory we present a simulation study, where we compare our test with the test of Grambsch and Therneau and the test of Gill and Schumacher.

Title: ***Cox regression and causal inference***
Speaker: **Torben Martinussen**, University of Copenhagen, Denmark
Schedule: *Wednesday 13 June, 12:30–13:00*
Abstract: Causal inference techniques has become very important tools in modern data analysis. There has been much debate whether special issues arises when one is considering time-to-event data. Classical tools for such data have evolved around the hazard function with the Cox regression model being the primary example. However, concerns has been raised whether this is a suitable function to consider when trying to draw causal inference, and one even speaks of "the hazards of hazards". In this talk I will try to shed some light on these issues.

Title: ***Efficiency in survival analysis: Towards a multivariate theory under censoring***
Speaker: **Arnold Janssen**, Heinrich Heine University, Germany
Schedule: *Wednesday 13 June, 13:00–14:00*
Abstract: The first part of the talk reviews hazard based efficiency results for univariate data under right random censoring:

1. Adaptive survival tests of log-rank type.
2. Efficient estimation of survival functionals.

The multivariate analysis starts with efficient testing of the null hypothesis of independence. A special example is Spearman's rank correlation test under censoring. Already at dimension two the estimation of the two dimensional survival function is a difficult affair, see Dabrowska, Gill, Stute or van der Laan. We offer an efficient point estimator of the multivariate survival function for a fixed point. They apply to estimators for hazard exponent measures of dependence.

Invited Session IS-43 • Thursday 14, 08:30-10:30 • Room Vietri

FUNCTIONAL DATA ANALYSIS II

Organizer(s): Aurore Delaigle (University of Melbourne)
Chair: Aurore Delaigle (University of Melbourne)

Title: ***A Mahalanobis-type distance for functional data***
Speaker: **Antonio Cuevas**, Universidad Autónoma de Madrid, Spain
Co-author(s): José R. Berrendero, Beatriz Bueno-Larraz
Schedule: *Thursday 14 June, 08:30–09:00*
Abstract: A new proposal for a distance between functional data is given. This new definition is inspired by that of the classical Mahalanobis distance for multivariate data. The machinery of Reproducing Kernel Hilbert Spaces is used to motivate the proposal. Some interesting properties of this new distance (concerning invariance, consistent estimation, sampling distribution in the Gaussian case and asymptotic distribution) are obtained. Some practical applications to exploratory data analysis and inference problems are discussed.

Title: ***Metric depth***
Speaker: **Alicia Nieto Reyes**, Universidad de Cantabria, Spain
Co-author(s): Heather Battey
Schedule: *Thursday 14 June, 09:00–09:30*
Abstract: For some function spaces there is more than one natural metric. However the choice of metric has important implications. Our proposed notion of depth takes the metric into consideration and it is atypical in that it accommodates functional data over any continuum as well as multivariate functional data. An illustration of the method on data from a study of Alzheimers patients is provided.

Title: ***Analyzing functional data over complex multi-dimensional domains***
Speaker: **Laura M. Sangalli**, Politecnico di Milano, Italy
Co-author(s): Luca Negri, Eleonora Arnone
Schedule: *Thursday 14 June, 09:30–10:00*
Abstract: We will present a novel class of models for the analysis of functional data defined over complex multidimensional domains, including curved bi-dimensional domains and complex three-dimensional domains. This class includes smoothing methods, regression methods and principal component analysis methods. These are implemented using numerical techniques such as finite elements and they are based on the idea of differential regularizations. An illustration to the analysis of neuroimaging data is provided. In this applicative domain, the proposed methods offer important advantages with respect to the best state of art techniques, allowing to correctly take into account to complex anatomy of the brain.

Title: ***Nonparametric mean estimation for big-but-biased data***
Speaker: **Ricardo Cao**, Universidade da Coruña (Spain); CITIC (Spain); ITMATI (Spain)

Co-author(s): Laura Borrajo
Schedule: *Thursday 14 June, 10:00–10:30*
Abstract: Some authors have recently warned about the risks of the sentence "with enough data, the numbers speak for themselves". Several problems coming from ignoring sampling bias in big data statistical analysis have been recently reported. The problem of nonparametric estimation of the mean in big data under the presence of sampling bias is considered in this work. This problem is studied when the biasing weight function is known (unrealistic) as well as for unknown weight functions (realistic). Two different scenarios are considered to remedy the problem of ignoring the weight function: (i) having a small sized simple random sample of the real population and (ii) having observed a sample from a doubly biased distribution. In both scenarios the problem is related to nonparametric density estimation, so kernel methods are used as auxiliary tools. Asymptotic expressions for the mean squared error of the estimators proposed are derived. This leads to asymptotic formulas for the optimal smoothing parameters involved. Some simulations are carried out to illustrate the performance of the nonparametric methods proposed. These methods are also applied to a dataset related to delay times in airlines.

Invited Session IS-44 • Thursday 14, 08:30-10:30 • Room Ravello

LOCALLY STATIONARY PROCESSES AND MULTIVARIATE TIME SERIES

Organizer(s): Wilfredo Palma & Giovanni Motta (Pontificia Universidad Catolica de Chile)

Chair: Wilfredo Palma (Pontificia Universidad Catolica de Chile)

Title: ***Time-varying covariances and correlations: Adaptivity Versus Positivity***
Speaker: **Giovanni Motta**, Pontificia Universidad Catolica de Chile, Chile
Schedule: *Thursday 14 June, 08:30–09:00*
Abstract: In this work we introduce a novel approach for estimating time-varying covariance matrices. Our estimators are based on nonlinear transformations of localized covariances and correlations. This new approach allows for different smoothing bandwidths while preserving positive definiteness.

Our approach is particularly attractive for locally stationary time series. More specifically, we address the estimation of two types of time-varying matrices: the contemporaneous covariance of

multivariate time series, and the auto-correlation of univariate time series.

Title: ***Towards a General Theory for Non-Linear Locally Stationary Processes***
Speaker: **Rainer Dahlhaus**, Heidelberg University, Germany
Co-author(s): Stefan Richter, Wei Biao Wu
Schedule: *Thursday 14 June, 09:00–09:30*
Abstract: In this paper 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.

Title: ***Detecting Frequency Changes in Time Series***
Speaker: **David Stoffer**, University of Pittsburgh, United States of America
Schedule: *Thursday 14 June, 09:30–10:00*
Abstract: This is preliminary work on detecting changes in the spectral density of a time series using simple nonparameteric or semiparametric methods. The problem is motivated by the detection of changes to the frequency of El Nino events sometime in the 1980s (which seems to be taken for granted by climatologists). We use the Whittle likelihood coupled with minimum description length to examine the Southern Osciallation Index. There are a number of other monitors of sea surface temperature, which makes the problem multivariate. Previous time domain methods such as fitting local AR models, which do not capture slight changes in frequencies. Vector AR models may be even more problematic. Since we are using the Whittle likelihood, the approach should extend easily to the multivariate case (we hope). I will discuss our preliminary methods.

Title: ***Stationary Subspace Analysis & Brain-Computer Interface***
Speaker: **Mohsen Pourahmadi**, Texas A&M University, United States of America
Schedule: *Thursday 14 June, 10:00–10:30*
Abstract: Stationary subspace analysis (SSA) is a recent time domain technique

for classification of brain-computer interface nonstationary data. It finds linear transformations of nonstationary multivariate processes which are stationary in the limited sense that the first two moments or means and covariances are time-invariant. The key optimization problem is that of finding a matrix minimizing the Kullback-Leibler divergence between Gaussian distributions measuring the non-constancy of the means and covariances across several segments. We present a frequency domain alternative to SSA for general multivariate second-order nonstationary processes. Using the asymptotic uncorrelatedness of the discrete Fourier transform of a stationary time series, a measure of departure from stationarity is introduced and minimized to find the stationary subspace. The dimension of the subspace, the key parameter, is estimated using a sequential testing procedure and its asymptotic properties are studied. We illustrate the broader applicability and better performance of the frequency domain method in comparison to time domain SSA methods through simulations and an application in analyzing EEG data from brain-computer interface experiments.

Invited Session IS-45 • Thursday 14, 08:30-10:30 • Room Furore

NEW TRENDS AND ADVANCES IN BAYESIAN NONPARAMETRICS

Organizer(s): Antonio Lijoi (Bocconi University)

Chair: Antonio Lijoi (Bocconi University)

Title: ***A Bayesian nonparametric spiked process prior for dynamic model selection***

Speaker: **Michele Guindani**, University of California, Irvine, United States of America

Schedule: *Thursday 14 June, 08:30–09:00*

Abstract: In many applications, investigators consider processes that vary in space and time, with the goal of identifying temporally persistent and spatially localized departures of those processes from a baseline or "normal" behavior. We propose a Bayesian nonparametric model selection approach for the analysis of spatio-temporal data, which takes into account the non-exchangeable nature of measurements collected over time and space. More specifically, a zero-inflated conditionally identically distributed (CID) species sampling prior is used to model temporal dependence in the selection, by borrowing information across time and assigning data to clusters associated

to either a null or an alternate process. Spatial dependences are accounted for by means of a Markov random field (MRF) prior, which allows informing the selection based on inferences conducted at nearby locations. We investigate the performances of our model by means of a simulation study and an application to a disease surveillance problem, for detecting outbreaks of pneumonia and influenza (P&I) mortality in the continental United States. We show how the proposed modeling framework compares favorably with respect to commonly adopted threshold methods for detecting outbreaks over time and also to recent proposals modeling more complex Markov switching dependences.

Title: ***Dynamic and distributed inference of latent geometric structures arising in topic modeling***

Speaker: **Long Nguyen**, University of Michigan, United States of America

Schedule: *Thursday 14 June, 09:00–09:30*

Abstract: We present a new model and algorithms for learning about the temporal dynamics of topic polytopes and related geometric objects that arise in topic model based inference. Our model is Bayesian nonparametric and the corresponding inference algorithm is able to discover new topics as the time progresses. We establish the connection between the modeling of topic polytope evolution, Beta-Bernoulli process and the Hungarian matching algorithm. Our method is shown to be several orders of magnitude faster than existing scalable topic modeling approaches, as demonstrated by experiments working with several million documents in a dozen minutes.

Title: ***Markov chains for dependent nonparametric priors***

Speaker: **Ramses H Mena**, Universidad Nacional Autonoma de Mexico, Mexico

Schedule: *Thursday 14 June, 09:30–10:00*

Abstract: Most discrete nonparametric priors are driven by a structural discrete random variable. In this work we explore a novel construction of Markov chains which suits well to define time-dependent nonparametric priors. Unlike other constructions available in the literature, the proposed approach includes the time-dependence at a very basic structural level, enabling an appealing analytical and computational control. The proposal is explored in both, discrete and continuous time, as well as for some well-known choices nonparametric priors. An algorithm is proposed and some illustrations with simulated and real data is presented.

Title: ***Distribution theory for hierarchical processes***

Speaker: **Igor Pruenster**, Università Bocconi, Italy
Schedule: *Thursday 14 June, 10:00–10:30*
Abstract: Hierarchies of discrete probability measures represent popular nonparametric priors in several applications. This is due to two key properties: (i) they naturally represent multiple heterogeneous populations; (ii) they produce ties across populations, resulting in a shrinkage property often described as "sharing of information".
A distribution theory for hierarchical random measures that are generated via normalization, thus encompassing both the hierarchical Dirichlet and hierarchical Pitman–Yor processes is presented. These results provide a probabilistic characterization of the induced (partially exchangeable) partition structure, including the distribution and the asymptotics of the number of partition sets, and a complete posterior characterization. Moreover, they also serve as building blocks for new simulation algorithms (both marginal and conditional) for Bayesian inference. Also other dependent random measures (based on additive and nested constructions) will be discussed and comparisons drawn.

Invited Session IS-46 • Thursday 14, 08:30-10:30 • Room Positano

INFERENCE FOR HIGH-DIMENSIONAL MATRIX AND NETWORK DATA
Organizer(s): Marianna Pensky (University of Central Florida)
Chair: Marianna Pensky (University of Central Florida)

Title: ***Change point inference for time dependent random graphs***
Speaker: **George Michailidis**, University of Florida, United States of America
Schedule: *Thursday 14 June, 08:30–09:00*
Abstract: We consider a change point problem on a random graph model with a large number of edges, where each edge changes state across time according to its own Markov transition matrix and independently of every other edge. We estimate the change point and Markov transition matrices before and after the change point by the method of maximum likelihood. Convergence rate and asymptotic distributions of these estimators are derived. Asymptotic distributions of the change point estimator are different depending on the regime that the signal-to-noise ratio -high-medium-low- falls in. Since the regime is not a priori known, we develop a data driven adaptive procedure that provides valid confidence intervals for the change point. The model and the results are illustrated on roll-call data from the US Congress.

Title: ***Structured Matrix Estimation and Completion***
Speaker: **Olga Klopp**, ESSEC Business School, France
Co-author(s): Yu Lu, Alexandre B. Tsybakov, Harrison H. Zhou
Schedule: *Thursday 14 June, 09:00–09:30*
Abstract: We study the problem of matrix estimation and matrix completion under a general framework. This framework includes several important models as special cases such as the gaussian mixture model, mixed membership model, bi-clustering model and dictionary learning. We consider the optimal convergence rates in a minimax sense for estimation of the signal matrix under the Frobenius norm and under the spectral norm. As a consequence of our general result we obtain minimax optimal rates of convergence for various special models.

Title: ***Estimation and clustering in a Dynamic Stochastic Block Model***
Speaker: **Marianna Pensky**, University of Central Florida, United States of America
Co-author(s): Teng Zhang
Schedule: *Thursday 14 June, 09:30–10:00*
Abstract: We consider a Dynamic Stochastic Block Model (DSBM) under the assumptions that the connection probabilities, as functions of time, are smooth and that only few nodes can switch their class memberships between two consecutive time points. The objective is estimation of the tensor of connection probabilities and clustering of the nodes. In particular, in the context of the DSBM, we derive a penalized least squares estimator of the tensor of connection probabilities and show that it satisfies an oracle inequality and also attains minimax lower bounds for the risk. For the purpose of clustering, we estimate the edge probability tensor by a kernel-type procedure and extract the group memberships of the nodes by spectral clustering. The procedure is computationally viable, adaptive to the unknown smoothness of the functional connection probabilities, to the rate of membership switching and to the unknown number of clusters. In addition, it is accompanied by non-asymptotic guarantees for the precision of estimation and clustering.

Invited Session IS-47 • Thursday 14, 08:30-10:30 • Room Maiori

ADVANCES IN OBJECT DATA ANALYSIS II
Organizer(s): Vic Patrangenaru (Florida State University)
Chair: Andrew Wood (University of Nottingham)

Title: ***Extracting multiscale shape information from multivariate and object data, with application to classification***

Speaker: **Wolfgang Polonik**, University of California, Davis, United States of America

Co-author(s): Gabriel Chandler

Schedule: *Thursday 14 June, 08:30–09:00*

Abstract: Guided by geometric considerations, we develop a novel method for extracting qualitative features of multivariate and object data, such as functional data. Similar to the famous kernel trick, we map data into a feature space, which consists of functions. In contrast to the kernel trick, our real-valued feature functions are defined on the interval [0,1], and thus can be plotted. By design, these functions contain information on different scales. For multivariate data, the two extreme scales correspond to density and data depth, respectively. The usefulness of this approach for statistical inference is illustrated by an application to classification. We also present theory to show that our procedure does not suffer from the curse of dimensionality.

Title: ***Principal Nested Submanifolds***

Speaker: **J. S. Steve Marron**, UNV, United States of America

Schedule: *Thursday 14 June, 09:00–09:30*

Abstract: In the age of Big Data a less well understand, but perhaps greater challenge is the analysis of Complex Data. Many different data types, including various shape representations and other types of imaging based data, are most effectively analyzed in terms of points on a manifold. An important approach to data visualization is Principal Component Analysis (PCA). Various existing methods, including backwards approaches, and Barycentric Subspaces, are seen to be special cases of a general framework called Principal Nested Submanifolds. This framework is seen to present new optimization challenges and open problems.

Title: ***A CLT for the intrinsic Frechet mean on compact Riemannian manifolds***

Speaker: **Andrew Wood**, University of Nottingham, United Kingdom

Schedule: *Thursday 14 June, 09:30–10:00*

Abstract: In a recent paper, Huckemann and Hotz (2015, Annals of the Institute of Statistics Mathematics) investigated the behaviour of the intrinsic Frechet mean on the unit circle. The results they obtained were rather nonstandard and surprising. In my talk I will describe a central limit theorem (CLT) for the intrinsic Frechet mean on a general compact symmetric Riemannian manifold that correctly accommodates the Huckemann and Hotz results on the circle. Moreover, this CLT

clarifies when nonstandard behaviour does and does not occur. The problem of consistently estimating the asymptotic covariance matrix in the class of nonstandard cases will also be discussed briefly. This talk is based on joint work with Thomas Hotz (Ilmenau) and Huiling Le (Nottingham)..

Invited Session IS-48 • Thursday 14, 08:30-10:30 • Room Amalfi

ALGEBRAIC METHODS AND NONPARAMETRIC STATISTICS

Organizer(s): Roberto Fontana & Fabio Rapallo (Politecnico di Torino & Università del Piemonte Orientale)

Chair: Roberto Fontana (Politecnico di Torino)

Title: ***Exact tests to compare contingency tables under quasi-independence and quasi-symmetry***

Speaker: **Fabio Rapallo**, Universita' del Piemonte Orientale, Italy

Schedule: *Thursday 14 June, 08:30–09:00*

Abstract: In the framework of contingency tables analysis, several square tables are considered under the quasi-independence (qi) or the quasi-symmetry (qs) model. Working within the class of log-linear models, a suitable model is defined and a goodness-of-fit test is introduced in order to verify if two or more tables fit a common qi (or qs) model against the alternative that each table follows a different qi (or qs) model. Such an exact test is based on classical tools of Algebraic Statistics, i.e., the computation of a Markov basis and a MCMC algorithm. When the comparison is limited to two tables, the Markov bases are characterized theoretically, while for larger comparisons the computations are performed through symbolic software. Applications to social mobility tables and rater agreement problems are discussed.

Title: ***Speeding up algebraic-based sampling via permutations***

Speaker: **Francesca Romana Crucinio**, University of Warwick, United Kingdom

Co-author(s): Roberto Fontana

Schedule: *Thursday 14 June, 09:00–09:30*

Abstract: Consider two samples of size n_1 and n_2 coming from some non-negative discrete exponential family. There exists a uniformly most powerful unbiased procedure to test if the two samples come from the same distribution, performed conditionally on the sum of their entries t. The resulting sample space is the fiber $\mathscr{F}_{n_1+n_2,t}$ of vectors

of size $n_1 + n_2$ and entries adding up to t. Vectors can be sampled from this set by building a Markov Chain Monte Carlo procedure which exploits Markov basis to connect all the elements of $\mathscr{F}_{n_1+n_2,t}$. The fiber can be partitioned into orbits of permutations and the probability of sampling a vector $\mathbf{y}$ given its orbit $\pi_{\mathbf{y}}$ is uniform. As a consequence, it is possible to sample orbits $\pi \subseteq \mathscr{F}_t$ via MCMC with the appropriate Markov basis, and then $\mathbf{y} \in \pi$ via standard Monte Carlo. The two sampling procedures above result in two well-known estimators: the indicator function and the permutation cumulative distribution function. Both estimators are unbiased, but the one provided by the permutation approach has the lowest variance and the lowest mean absolute deviation. These theoretical results are verified by a simulation study, showing that the permutation approach grants convergence in the least steps too.

Title: ***Adaptive global thresholding on the sphere***
Speaker: **Claudio Durastanti**, Ruhr-Universität Bochum, Germany
Schedule: *Thursday 14 June, 09:30–10:00*
Abstract: This talk is concerned with the study of the adaptivity properties of nonparametric regression estimators over the d-dimensional sphere within the global thresholding framework. The aforementioned estimators are constructed by means of spherical wavelets, named needlets, which enjoy strong concentration properties in both harmonic and real domains. The related rates of the Lp risks are hence established, focussing on their minimax properties and proving their optimality over a scale of nonparametric regularity function spaces, namely, the Besov spaces.

Invited Session IS-49 • Thursday 14, 08:30-10:30 • Room Procida

RECENT ADVANCES IN EXACT INFERENCE
Organizer(s): Pralay Senchaudhuri (Cytel Inc.)
Chair: Pralay Senchaudhuri (Cytel Inc.)

Title: ***Comparing Saddle-point approximations and Exact methods in Logistic Regression***
Speaker: **Nitin Patel**, Cytel Inc., United States of America
Co-author(s): Pralay Senchaudhuri, Jaydeep Bhattecharyya
Schedule: *Thursday 14 June, 08:30–09:00*
Abstract: In logistic regression analysis, when sample sizes are small or the number of responses are few (rare events), the widely used standard

large sample normal approximations are unreliable. Exact Logistic Regression (ELR) does not rely on asymptotics and can be used to estimate distributions of Maximum Likelihood Estimates, perform tests of significance and compute confidence intervals. They can be used for joint, marginal and conditional estimation. However, they can be demanding in computational time and memory and can fail to produce useful results when the covariates are not lattice variables. In this talk we will report on computational experiments to compare higher order Saddle-Point Approximations (SPA) that overcome these limitations with ELR using simulated data as well as published datasets. We will also explore 'hybrid' methods, such as combining Monte Carlo sampling from ELR distributions with SPA, to leverage advantages of both methods.

Title: ***Exact Meta-Analysis Using a Permutation-Based Approach***
Speaker: **Chris Corcoran**, Utah State University, United States of America
Co-author(s): Brinley Zabriskie, Pralay Senchaudhuri
Schedule: *Thursday 14 June, 09:00–09:30*
Abstract: The increasingly widespread use of meta-analysis has led to growing interest in meta-analytic methods for rare events and sparse data – conventional approaches tend to perform very poorly in such settings. Recent work in this area has provided options for sparse data, but these are still often hampered when heterogeneity across the available studies differs according to the exposure or treatment group. We propose a permutation-based approach that accommodates this common contingency, providing more reliable statistical tests when such patterns of heterogeneity are observed. Our approach also can be used as a basis for computing exact confidence intervals when estimating effect size.

Title: ***Improved Power of Familywise Error Rate Procedures for Discrete Data under Dependency***
Speaker: **Li He**, Merck Research Laboratories, United States of America
Co-author(s): Joseph Heyse
Schedule: *Thursday 14 June, 09:30–10:00*
Abstract: In many applications where it is necessary to test multiple hypotheses simultaneously, the data encountered are discrete. In such cases, it is important for multiplicity adjustment to take into account the discreteness of the distributions of the p-values, to assure that the procedure is not overly conservative. To this end, we derive discrete versions of some commonly used multiple testing procedures that control the familywise error rate, the probability of making any false rejection. Taking advantage of the fact that

the exact permutation or exact pairwise permutation distributions of the p-values can often be determined when the sample size is small, we also derive procedures that incorporate the dependence structure through the exact permutation or exact pairwise permutation distributions. The performance of the proposed procedures is investigated through simulation studies and two applications. The results show that by incorporating both discreteness and dependency of p-value distributions, gain in power can be achieved.

Title: ***Likelihood penalties for exact conditional logistic regression***
Speaker: **Georg Heinze**, Medical University of Vienna, Austria
Schedule: *Thursday 14 June, 10:00–10:30*
Abstract: Exact conditional logistic regression (ECLR) estimates regression coefficients (log odds ratios) by conditioning on the sufficient statistics of nuisance covariates. ECLR became feasible when efficient algorithms to compute these conditional permutational distributions became available. While ECLR enables exact inference about regression coefficients, maximizing the conditional likelihood does not necessarily produce more accurate point estimates than unconditional multivariable logistic regression (UMLR). We propose to consider likelihood penalties, e.g., the Firth penalty which removes the leading term of the bias expansion of regression coefficients. We also show how ridge or lasso penalties are easily implemented. Penalties can be interpreted as Bayesian prior distributions centered around zero; Firth, ridge and lasso penalties correspond to Jeffreys, normal and double exponential priors, respectively. Recently, the logF(m,m) prior was proposed which in UMLR is easily incorporated by data augmentation and also has a frequentist interpretation as likelihood penalty. All these priors/penalties provide shrinkage and can reduce the mean squared error of the point estimates. While for predictive purposes the amount of penalty is often determined by cross-validation, classical inference demands to fix the penalty in advance. We will propose some default penalties and will compare their performances in analyses of real and simulated data.

Plenary Session PL-03 • Thursday 14, 11:00-12:00 • Auditorium Tafuri

PREDICTION WITH CONFIDENCE – A GENERAL FRAMEWORK FOR PREDICTIVE INFERENCE

by Regina Y. Liu
Rutgers University, USA

Chair: Dimitris Politis (UCSD)

Abstract: We propose a general framework for prediction in which a prediction is in the form of a distribution function, called 'predictive distribution function'. This predictive distribution function is well suited to prescribing the notion of confidence under the frequentist interpretation, and it can provide meaningful answers for prediction-related questions. A general approach under this framework is formulated and illustrated using the so-called confidence distributions (CDs). This CD-based prediction approach inherits many desirable properties of CD, including its capacity to serve as a common platform for directly connecting the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. We discuss the theory underlying the CD-based predictive distribution and related efficiency and optimality issues. We also propose a simple yet broadly applicable Monte-Carlo algorithm for implementing the proposed approach. This concrete algorithm together with the proposed definition and associate theoretical development provide a comprehensive statistical inference framework for prediction. Finally, the approach is demonstrated by simulation studies and a real project on predicting the incoming volume of application submissions to a government agency. The latter shows the applicability of the proposed approach to dependent data settings as well.
This is joint work with Jieli Shen, Goldman Sachs, and Minge Xie, Rutgers University.

Contributed Session CS-20 • Thursday 14, 12:00-13:00 • Room Ravello

IMPERFECTLY OBSERVED DATA II
Chair: Stanislav Nagy (Charles University, Prague)

Title: ***Estimation of conditional prevalence from group testing data with***

missing covariates

Speaker: **Wei Huang**, University of Melbourne, Australia
Co-author(s): Aurore Delaigle, Shaoke Lei
Schedule: *Thursday 14 June, 12:00–12:20*
Abstract: We consider estimating the conditional prevalence of a disease from data pooled according to the group testing mechanism. Consistent estimators have been proposed in the literature, but they rely on the data being available for all individuals. In infectious disease studies where group testing is frequently applied, the covariate is often missing for some individuals. There, unless the missing mechanism occurs completely at random, applying the existing techniques to the complete cases without adjusting for missingness does not generally provide consistent estimators, and finding appropriate modifications is challenging. We develop a consistent adjusted spline estimator, derive its theoretical properties, and show how to adapt local polynomial and likelihood estimators to the missing data problem. We illustrate the numerical performance of our methods on simulated and real examples.

Title: ***Wavelet estimation in multiplicative regression***
Speaker: **Fabien Navarro**, CREST, ENSAI, France
Co-author(s): Christophe Chesneau, Junke Kou
Schedule: *Thursday 14 June, 12:20–12:40*
Abstract: We consider the estimation of an unknown regression function from a nonparametric regression model with a multiplicative noise. A linear wavelet estimator is developed. Using standard techniques, we evaluate its theoretical performances under the Lp risk. A numerical study illustrates the performances of the method.

Title: ***Nonparametric analysis of the shape of random functions***
Speaker: **Stanislav Nagy**, Charles University, Czech Republic
Schedule: *Thursday 14 June, 12:40–13:00*
Abstract: In many situations, the shape of functional data is an important feature that must be taken into account in statistical analysis. The information about the shape can be extracted from the derivatives of the sample trajectories. Though, this approach can be applied only if the curves are regular and smooth, and the derivatives must be estimated. We present a simple alternative to this methodology based on the notion of data depth. This technique does not require smoothness or continuity of the curves, yet provides fine recognition of the shape traits of the data. Possible applications include the analysis of (multivariate) functional and image data.

NONPARAMETRIC INFERENCE AND ESTIMATION FOR CORRELATED DATA

Chair: Josu Arteche (University of the Basque Country UPV/EHU)

Title: ***About Kendall's regression***

Speaker: **Jean-David Fermanian**, Crest-ENSAE, France

Co-author(s): Alexis Derumigny

Schedule: *Thursday 14 June, 12:00–12:20*

Abstract: Conditional Kendall's tau is a measure of dependence between two random variables, conditionally on some observed covariates. We propose and study nonparametric estimators of such quantities using kernel smoothing techniques. Then, we assume a regression-type relation between conditional Kendall's tau and covariates, in a parametric setting with possibly a large number of regressors. This model may be sparse, and the underlying parameter is estimated through a penalized criterion. The theoretical properties of all the estimators are stated. We prove non-asymptotic bounds with explicit constants that hold with high probability. We derive their consistency, their asymptotic law and some oracle properties. Some simulations evaluate the relevance of such procedures.

Title: ***Locally adaptive and wavelet regression for compositional data***

Speaker: **Andrej Srakar**, Institute for Economic Research (IER), Ljubljana and Faculty of Economics, University of Ljubljana, Slovenia

Schedule: *Thursday 14 June, 12:20–12:40*

Abstract: Regression for compositional data has been so far largely considered only from a parametric point of view. Aitchison (1982) and Hijazi and Jernigan (2009) modelled regression of a compositional response on a real predictor assuming, as distribution for residuals, the Dirichlet or the logistic-normal distributions. For the same problem, TolosanaDelgado and Van Den Boogart (2011) and Egozcue et al. (2012) proposed a linear model using ilr coordinates of the response, allowing ordinary least squares theory on the space of coordinates. In a recent article, Di Marzio, Panziera and Venieri (2015) extended this to nonparametric situations, introducing local constant and local linear smoothing for regression with compositional data and treating the cases when either the response, the predictor or both of them are compositions. In our analysis, we extend their analysis to locally adaptive estimators, in particular Haar wavelets. We present a detailed statistical and mathematical analysis, some comparison

(simulation) results with the performance of other existing estimators for regression with compositional data, while, finally, applying the results to two case studies from economics (inference for inequality indices/data).

Title: ***Exact Local Whittle estimation in general long memory series with multiple poles***

Speaker: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain

Schedule: *Thursday 14 June, 12:40–13:00*

Abstract: A semiparametric estimation technique is proposed for jointly estimating all the memory parameters in general long memory time series with a persistent cyclical behaviour. Consistency and asymptotic normality are proved for stationary, non-stationary and non-invertible series, allowing for straightforward standard inference of interesting hypotheses such as the existence of unit roots or equality of memory parameters at some or all seasonal frequencies, which can be used as a prior test for the application of seasonal differencing filters. The finite sample performance is analysed in an extensive Monte Carlo and an application to a U.S. Industrial Production index.

Contributed Session CS-22 • Thursday 14, 12:00-13:00 • Room Positano

LARGE DATA CLASSIFICATION

Chair: Yakov Yu. Nikitin (Saint-Petersburg State University)

Title: ***Algorithm for automatic description of historical series of forecast error in electrical power grid***

Speaker: **Gaia Ceresa**, RSE, Italy

Co-author(s): Andrea Pitto, Diego Cirio, Nicolas Omont

Schedule: *Thursday 14 June, 12:00–12:20*

Abstract: The EU-FP7 iTesla-project developed a toolbox that assess dynamic security of large electric power systems in presence of forecast uncertainties. In particular, one module extracts plausible realizations of the stochastic variables (power injections of RES –Renewable Energy Sources, load power absorptions). It is built upon historical data series of hourly forecasts and realizations of the stochastic variables at HV (High Voltage) nodes in the French transmission grid. Data reveal a large diversity of forecast error distributions: characterising them allows to adapt the module to the data, improving the results. The algorithm here presented is aimed to

automatically classify all the forecast error variables and to cluster them into "smoother" variables. The main steps of the algorithm are: filtering of the variables with too many missing data or too low variance; outliers detection by two methods (Chebyshev inequality, quantile method); separation of unimodal variables from multimodal ones by exploiting a peak counting algorithm, Gaussian mixtures, comparison with asymmetrical distributions, multimodality index; clustering of the multimodal variables whose sum is unimodal, comparing two alternative algorithms (the former based on hierarchical clusterization, accounting for correlation and geographical closeness, and the latter on the identification of the same initial characters in the identification codes).

Title: ***Disease classification based on big data microbiome analytics***

Speaker: **Martina Zhelcheva Mincheva**, Temple University, United States of America

Co-author(s): Jun Chen, Hongzhe Li

Schedule: *Thursday 14 June, 12:20–12:40*

Abstract: Next generation sequencing technologies make it possible to survey microbial communities by sequencing nucleic acid material extracted from multiple samples. This paper proposes a method to test the hypothesis that two sets of samples have the same microbial composition. Its generality stems from the fact that the test is nonparametric and requires no assumptions on the probability distributions. It is also applicable for varying set sizes and sample sizes. Its computational efficiency comes from the exact asymptotic distribution of the proposed test statistic. Extensive data analysis shows that the test is significantly faster than the permutation-based multivariate analysis of variance using distance matrices (permanova) (McArdle & Anderson, 2001). At the same time, it has correct type 1 errors and comparable power, which makes it preferable in the analysis of large-scale microbiome data.

Title: ***Local Bahadur efficiencies of scale-free tests of normality based on a recent characterization***

Speaker: **Yakov Yu. Nikitin**, Saint-Petersburg State University, Russian Federation,; National Research University - Higher School of Economics, St.Petersburg, Russia.

Schedule: *Thursday 14 June, 12:40–13:00*

Abstract: We develop two scale-free tests of normality based on the recent characterization of the symmetric normal law by Ahsanullah, Kibria and Shakil (2014). Both test statistics have an U-empirical structure, but the first one is of integral type, while the second one is of

Kolmogorov type. We study the limiting behavior and logarithmic large deviations of these test statistics under the null-hypothesis. Next, we calculate their local exact Bahadur efficiencies for location, skew and contamination alternatives via the upper bound for exact slopes in terms of Kullback-Leibler information. These efficiencies in many cases are rather high. For instance, we obtain the value 0.977 for the integral test in the case of location and skew alternatives. In our opinion, both tests can be added to the existing set of normality tests due to their high efficiency properties and relative simplicity. This research was supported by the grant of RFBR 16-01-00258 and by the grant SPbGU - DFG 6.65.37.2017.

Contributed Session CS-23 • Thursday 14, 12:00-13:00 • Room Procida

FUNCTIONAL DATA I

Chair: Silvia Novo Díaz (Universidade da Coruña)

Title: ***A Functional Version of the GARCH(p,q) model***
Speaker: **Sebastian Kühnert**, The University of Rostock, Germany
Schedule: *Thursday 14 June, 12:00–12:20*
Abstract: The GARCH(p,q) model with integers p,q is usually applied in order to model real valued financial time series and will be established for time series which take their values in function spaces of measurable functions on the domain [0,1].
Necessary and sufficient conditions for the existence of strictly stationary solutions will be provided for $L^p[0,1]$ spaces and Banach spaces endowed with the supremum norm.
Estimators are constructed for the shift term and the coefficient operators of the functional GARCH(p,q) equation. For these estimators asymptotic upper bounds are deduced in the sense of convergence in probability.

Title: ***A test of exogeneity in the functional linear regression model***
Speaker: **Manuela Dorn**, University of Bayreuth, Germany
Co-author(s): Melanie Birke, Carsten Jentsch
Schedule: *Thursday 14 June, 12:20–12:40*
Abstract: For the functional linear regression model

$$Y = \int_0^1 \beta(t)X(t)dt + \sigma U$$

a common question is whether the functional regressor X is endogenous, i.e. correlated with the error term U. Considering the case that the unknown slope parameter β is an element of the Sobolev Space of periodic functions and assuming second order stationarity of the regressor, a test of exogeneity is proposed. Using functional instrumental variables, the test statistic is constructed from two estimators that are consistent under exogeneity but have different behavior under endogeneity. As the estimation of β is an ill-posed inverse problem, thresholding is used for regularization. However, as it is not applicable in the functional context, the test statistic must have a different form than the one used in the original Hausman test. The asymptotic distribution of the test statistic is derived and a bootstrap method is proposed. Finally, the finite-sample performance of the test is checked by a small simulation study.

Title: ***Data-driven kNN estimation in functional single-index regression***

Speaker: **Silvia Novo Díaz**, Grupo de Investigación MODES, Departamento de Matemáticas, Universidade da Coruña, Spain; Centro de Investigación de Tecnoloxías da Información e da Comunicación (CITIC)

Co-author(s): Germán Aneiros Pérez, Philippe Vieu

Schedule: *Thursday 14 June, 12:40–13:00*

Abstract: In the functional data framework, regression problems involve infinite-dimensional variables and the dimensionality problem becomes very important. In this context, the literature is mainly concentrated either around nonparametric models or around linear model, but semiparametrics is still a very underdeveloped field. Accordingly, a study of a functional semiparametric model, namely the Functional Single-Index Model (FSIM), is presented for which a new estimation procedure based on kNN ideas is developed. General asymptotic results for the kNN procedure are stated, with main interest of being uniform over all the parameters of the model. This uniformity feature allows to derive immediately results for random data-driven choices of these parameters, making the procedure directly applicable in practice. Similar results for the standard kernel approach are derived too. An application to real data is included which illustrates how the kNN approach outperforms standard procedures. It also shows how the semiparametric feature of the FSIM has not only nice predictive performance but it also provides easily interpretable and representable outputs.

Special Invited Session SIS-06 • Thursday 14, 12:00-13:00 • Auditorium Tafuri

INFORMATION MATRICES AND GROUP STRUCTURES IN SUFFICIENT DIMENSION REDUCTION

by Francesca Chiaromonte[1] & Debmalya Nandy[2] & Weixin Yao[3] & Bruce Lindsay[2] & Yang Liu[4] & Bing Li[2]

[1]Pennsylvania State University, USA and Sant'Anna School of Advanced Studies, Italy; [2]Pennsylvania State University, USA; [3]UC Riverside, USA; [4]Bank of America, USA

Speaker: Francesca Chiaromonte
Chair: Piercesare Secchi (Politecnico di Milano)

Abstract: In most contemporary scientific fields, regression and classification problems comprise a large number of potential predictor variables. With a literature dating back to the early '90s, Sufficient Dimension Reduction (SDR) provides a nimble and effective toolkit for handling these high dimensional supervised problems. Specifically, SDR techniques identify a small number of composite features, most commonly linear combinations, which capture the core of the relationship between the response and the original predictor variables. In this talk we will highlight two important trajectories along which SDR has evolved in recent years, namely: (i) methodology based on non-parametric estimation of conditional mean or density functions – which eliminates restrictive assumptions typical of earlier reduction techniques; and (ii) approaches that leverage known structures (e.g. groups) to create more meaningful reduced representations of the data. As instances, we will focus on two novel tools: the Covariate Information Matrix (CIM) and the structured Ordinary Least Squares (sOLS). Work on CIM is joined with Debmalya Nandy (PSU), Weixin Yao (UC Riverside) and the late Bruce Lindsay (PSU). Work on sOLS is joined with Yang Liu (Bank of America) and Bing Li (PSU).

Invited Session IS-50 • Thursday 14, 14:30-16:30 • Room Vietri

GRAPHS AND TREES IN OBJECT ORIENTED DATA ANALYSIS

Organizer(s): James Stephen Marron (UNC)
Chair: James Stephen Marron (UNC)

Title: ***Beyond trees: Stratified spaces for variable topology data***

Speaker: **Aasa Feragen**, University of Copenhagen, Denmark

Schedule: *Thursday 14 June, 14:30–15:00*

Abstract: In the past 10 years, spaces of trees have been a popular and exotic playground for geometric statisticians: These spaces are not smooth, but stratified, and give rise to surprising phenomena that were not known from the more traditional manifold statistics. Computation in spaces of trees often requires both geometry and combinatorics, and tree-spaces have formed a meeting spot for many different branches of mathematics and computational sciences. While trees are interesting in their own right, there are larger families of objects, or data types, closely related to them, which have not yet received quite as much attention: Point sets and graphs. In this talk, I will discuss

- Why point sets are graphs are interesting in their own right, from an application perspective
- How point sets and graphs are modelled naturally as residing in stratified spaces
- What we already know about point sets and graphs, highlighting frequently overlooked related work (not by me)
- Open problems in modelling, statistics and computation, highlighting in particular the role of labels and their combinatorics in tree-space computation and what that means for point sets and graphs.

Title: ***Estimation and testing of functions of phylogeny and abundance***

Speaker: **Amy Willis**, Department of Biostatistics, University of Washington, United States of America

Schedule: *Thursday 14 June, 15:00–15:30*

Abstract: In many biological settings, estimating a function of the phylogenetic tree is more important than estimating the tree itself. We present new work on estimating summary statistics of phylogenetic trees, with a special focus on variance estimation and hypothesis testing. We particularly focus on estimating phylogenetically-weighted alpha and beta diversity, which summarise ecological communities with respect to both the phylogeny and relative abundance of the taxa. At the core of our method is the insight that neither the tree nor the relative abundance are known, and both are estimated with error. We show that propagating variance from both sources is integral to valid testing. We use the methodology to examine the extent to which microorganisms are shared between cohabiting human hosts.

Title: ***Mean and Variance for Tree-Shaped Data: Computation and***

Applications

Speaker: **Megan Owen**, Lehman College, City University of New York (CUNY), United States of America

Schedule: *Thursday 14 June, 15:30–16:00*

Abstract: Data generated in such areas as evolutionary biology and medical imaging are frequently tree-shaped, and thus non-Euclidean in nature. As a result, standard techniques for analyzing data in Euclidean spaces become inappropriate, and new methods must be used. One such framework is the space of metric trees constructed by Billera, Holmes, and Vogtmann. This space is a non-positively curved, or CAT(0), polyhedral cone complex, with a unique geodesic (shortest path) between any two trees and a well-defined notion of a mean tree. We give an algorithm for computing the Frechet mean that is fixed-parameter tractable in the number of different edges in the set of input trees, as well as give some applications of the mean and variance to phylogenetics.

Title: ***Modelling data in tree space via transition kernels of Brownian motion***

Speaker: **Tom M. W Nye**, Newcastle University, United Kingdom

Schedule: *Thursday 14 June, 16:00–16:30*

Abstract: Data sets consisting of samples of evolutionary trees, or phylogenies, for some fixed set of species arise in many different contexts in evolutionary biology. Analysing such data sets is challenging, since the space of all phylogenetic trees is highly non-Euclidean, although it has a natural structure as a geodesic metric space. In particular, it is very difficult to construct tractable distributions on tree space, so most analyses have relied on minimizing sums of squared geodesic distances. Here we propose an alternative approach by constructing distributions via random walks in tree space and their continuous-time limits. The transition kernels of these stochastic processes can be used to construct probabilistic models on tree space. We demonstrate how these models share some properties of population genetic models of relationships between gene trees and species tree, and how to carry out inference via a bridge construction in tree space.

Invited Session IS-51 • Thursday 14, 14:30-16:30 • Room Ravello

HIGH DIMENSIONAL STATISTICAL INFERENCE

Organizer(s): Francesco Giordano & Maria Lucia Parrella (Unversity of Salerno)

Chair: Francesco Giordano (Unversity of Salerno)

Title: ***A differential-geometric approach to high-dimensional risk regression models***

Speaker: **Luigi Augugliaro**, Università degli Studi di Palermo, Italia

Co-author(s): Angelo Mineo

Schedule: *Thursday 14 June, 14:30–15:00*

Abstract: Clinical studies where patients are routinely screened for many genomic features are becoming more routine. In principle, this holds the promise of being able to find genomic signatures for a particular disease. In particular, cancer survival is thought to be closely linked to the genomic constitution of the tumour. Discovering such signatures will be useful in the diagnosis of the patient, may be used for treatment decisions and, perhaps, even the development of new treatments. However, genomic data are typically noisy and high-dimensional, not rarely outstripping the number of patients included in the study. Regularized survival models have been proposed to deal with such scenarios. These methods typically induce sparsity by means of a coincidental match of the geometry of the convex likelihood and (near) non-convex regularizer. The disadvantages of such methods are that they are typically non-invariant to scale changes of the covariates, they struggle with highly correlated covariates and they have a practical problem of determining the amount of regularization. We propose an extension of the differential geometric least angle regression method for sparse inference in relative risk regression models.

Title: ***Portfolio Selection via Hybrid Graphical Least Square Estimation***

Speaker: **Hongsheng Dai**, University of Essex, United Kingdom

Co-author(s): Saeed Al Dahmani, Qiaozhen Zhang

Schedule: *Thursday 14 June, 15:00–15:30*

Abstract: Regression analysis is often used in the portfolio optimization problem within the Markowitz mean-variance framework. The challenge exists when the number of assets v is larger than the sample size N. We consider a Graphical Least Squares method to deal with such problems. Unlike the regularization methods

such as Ridge regression, LASSO and LARS, which always give biased estimates, the proposed method can give unbiased estimates for some parameters. The new approach assists in improving the portfolio performance by increasing the portfolio's expected return and decreasing its risk, which consequently affects the Sharpe ratio. Another advantage of the proposed method is that it constructs a non-sparse (saturated) portfolio, which is more diversified in terms of stocks, and reduces the stock-specific risk.

Title: ***On limit horizons in high dimensional inference***
Speaker: **Soumendra N. Lahiri**, NCSU, United States of America
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: We consider a common situation arising in many high dimensional statistical inference problems where the dimension d diverges with the sample size n and the statistic of interest is given by a function of component-wise summary statistics. The limit distribution of the statistic of interest is often influenced by an intricate interplay of underlying dependence structure of the component-wise summary statistics. Here, we introduce a new concept, called limit horizon (L.H.) that gives the boundary of the growth rate of d as a function of n where the natural approach to deriving the limit law by iterated limits works. Further, for d growing at a faster rate beyond the L.H., the natural approach breaks down. We investigate the L.H. in some specific high dimensional problems.

Title: ***GRID: A method for variable and model selection in high dimensional nonparametric regression***
Speaker: **Francesco Giordano**, Dept of Economics and Statistics, Unversity of Salerno, Italy
Co-author(s): Soumendra Nath Lahiri, Maria Lucia Parrella
Schedule: *Thursday 14 June, 16:00–16:30*
Abstract: In the context of nonparametric regression in high dimensions a new method, called the GRID, is proposed when only a relatively small subset of a large number of variables are relevant and may have nonlinear effects on the response. We develop a technique for variable and model selection. Moreover, we make the estimation of the true low-dimensional regression function, allowing any degree of interactions among the relevant variables. The proposed method, combines empirical likelihood based marginal testing with the local linear estimation machinery in a novel way to select the relevant variables. Further, it provides a simple graphical tool for identifying the low dimensional nonlinear structure of the regression function.

NONPARAMETRIC DATA SCIENCE

Organizer(s): Subhadeep Mukhopadhyay (Temple University)
Chair: Subhadeep Mukhopadhyay (Temple University)

Title: ***Characteristic Tensor Product Kernels***
Speaker: **Zoltan Szabo**, Ecole Polytechnique, France
Co-author(s): Bharath Sriperumbudur
Schedule: *Thursday 14 June, 14:30–15:00*
Abstract: Hilbert-Schmidt independence criterion (HSIC) is among the most popular and efficient approaches in data science and statistics to measure the dependence of random variables. Thanks to its kernel-based foundations, HSIC is applicable on numerous domains; examples include documents, images, trees, graphs, time series, dynamical systems, sets or permutations. Despite its tremendous practical success, quite little is known about when HSIC characterizes independence. I am going to provide a complete answer to this question. [Preprint: `https://arxiv.org/abs/1708.08157`. Code: `https://bitbucket.org/szzoli/ite-in-python/`]

Title: ***Probabilistic Index Models for Flexible Rank Based Inference***
Speaker: **Olivier Thas**, Ghent University, Belgium; Hasselt University, Belgium; University of Wollongong, Australia
Co-author(s): Jan De Neve, Karel Vermeulen, Gustavo Amorin, Stijn Vansteelandt
Schedule: *Thursday 14 June, 15:00–15:30*
Abstract: Probabilistic Index Models (PIM) were introduced in Thas et al. (2012) as a class of semiparametric models for the conditional probabilistic index (PI) which is defined as the probability that $Y < Y^*$ given X and X^*, where X and X^* are the covariates corresponding to the outcomes Y and Y^*, respectively. De Neve and Thas (2015) demonstrated that the PIM methodology generates many of the classical rank tests for factorial designs (e.g. Wilcoxon-Mann-Whitney, Kruskal-Wallis), but the flexibility of the model allows to generate rank-type tests for many other designs, including correcting for continuous covariate effects. The original semiparametric theory resulted in asymptotically normal distributed parameter estimators. Current work has focused on the construction of semiparametric efficient estimators and on improved inference in small samples.
References:
De Neve, J. and Thas, O. (2015). A Regression Framework for

Rank Tests Based on the Probabilistic Index Model. Journal of the Americal Statistical Association, 110, 1276-1283.
Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012) Probabilistic Index Models (with discussion, read paper). Journal of the Royal Statistical Society, Series B, 74(4), 623-671.

Title: ***Survival Tree-Based Methods for Nonstandard Data***
Speaker: **Jeffrey S. Simonoff**, New York University, United States of America
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: Tree methods (recursive partitioning) are a popular class of nonparametric methods for analyzing data. One extension of the basic tree methodology is the survival tree, which applies recursive partitioning to censored survival data. This has mainly been designed for right-censored data. Survival tree methods can also be applied to two important but less standard data types, left-truncated and right-censored (LTRC) data and interval-censored data. Further, LTRC trees can be used to analyze survival data with time-varying covariates, essentially building a time-varying covariates survival tree. Implementation of the methods is easy, and simulations and real data analysis results show that the proposed methods work well from both a predictive point of view and in uncovering tree structure in the underlying survival process. Such methods also can be used to produce ensemble methods such as (random) survival forests that can result in more accurate predictions by reducing the natural sample-to-sample variability of tree methods through resampling. This is joint work with Halina Frydman, Wei Fu, and Weichi Yao.

Title: ***Development of A Universal Modelling Strategy for Conditional Distributions***
Speaker: **Subhadeep Mukhopadhyay**, Temple University, United States of America
Schedule: *Thursday 14 June, 16:00–16:30*
Abstract: A new representation theory of conditional distribution will be presented along with its unifying role for statistical data modelling that permits us to integrate a large number of diverse (Y,X) modelling tools (e.g., k-sample problem, generalized linear model, quantile regression, probability index model, composition data regression, density regression, classification) into a single general algorithmic framework. This surprising simplicity and unity of our new formalism will be demonstrated using wide-ranging real applications.

MODEL AVERAGING AND RELATED TOPICS

Organizer(s): Jeffrey Racine (McMaster University)
Chair: Jeffrey S. Racine (McMaster University)

Title: ***Criterion-Based Model Averaging***
Speaker: **Byunghoon David Kang**, Lancaster University, United Kingdom
Co-author(s): Seojeong Jay Lee
Schedule: *Thursday 14 June, 14:30–15:00*
Abstract: We study a new class of model averaging estimators. The proposed estimator is defined as the maximizer of the weighted average of criterion or objective functions under each of the models considered, instead of averaging the parameter estimates over the models. We consider likelihood-based criterion function as well as non-likelihood setups such as the least squares (LS) and the generalized method of moments (GMM).

Title: ***Optimal Estimation with Complete Subsets of Instruments***
Speaker: **Youngki Shin**, McMaster University, Canada
Co-author(s): Seojeong Lee
Schedule: *Thursday 14 June, 15:00–15:30*
Abstract: In this paper we propose a two-stage least squares estimator whose first stage estimator is based on the complete subset averaging. The size of the subsets to be averaged is chosen to minimize the approximate mean squared error criteria. We give a theoretical derivation of the criteria function and investigate its properties. We also show the asymptotic optimality of the subset size choice. In Monte Carlo simulation studies, we find that the proposed estimator outperforms alternative estimators in some designs especially when instruments are correlated and the strength of endogeneity is high.

Title: ***Latent Group Structure with Heterogenous Distribution: Identification and Estimation***
Speaker: **Wendun Wang**, Erasmus University Rotterdam, Tinbergen Institute
Co-author(s): Heng Chen, Xuan Leng
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: Panel data are often characterized by individual heterogeneity, and a popular way of modelling heterogeneity is to cluster units into groups. The group pattern of heterogeneity not only exists in the mean but also in the distribution of a group. To identify latent groups and recover heterogeneous distribution for identified

groups, we propose a clustering method based on composite quantile regressions. We show that combining multiple quantiles improves the efficiency of the group membership estimate when the group structure is common across quantiles. We also propose amenable tests for commonality of group structure across quantiles. Asymptotic theories for the proposed estimators are established while their finite sample performance are demonstrated by simulations. We apply the proposed methods to analyse cross-country output effect of infrastructure capital.

Invited Session IS-54 • Thursday 14, 14:30-16:30 • Room Maiori

NEW CHALLENGES AND SOLUTIONS TO COMPLEX DATA PROBLEMS WITH APPLICATIONS

Organizer(s): Ronghui Lily Xu (University of California, San Diego)
Chair: Jelena Bradic (University of California, San Diego)

Title: ***Revisiting Huber's M-Estimation: A Tuning-Free Approach***
Speaker: **Wenxin Zhou**, UCSD, United States of America
Co-author(s): Chao Zheng, Lili Wang, Wen Zhou
Schedule: *Thursday 14 June, 14:30–15:00*
Abstract: The robustification parameter, which balances bias and robustness, has played critical role in recent development on robust estimation for heavy-tailed data. Practically, it can be tuned by cross-validation, a general computation-based model fitting technique that is commonly used to predict the performance of a model on a validation set. The underpinning mathematical principle, however, is disregarded. Guided by its theoretically optimal expression and also inspired by the censored moment equation approach [Ann. Probab. 18, 1284–1341], we propose a new turning-free scheme to adaptively choose the robustification parameter from data. Starting with the basic problem of estimating the mean, we further extend the proposed method to regressions in both low and high dimensional settings. In a unified manner, the proposed method will be referred to as the data-adaptive Huber method. We illustrate its promising performance with extensive numerical experiments.

Title: ***Variable Selection for Highly Correlated Predictors***
Speaker: **Annie Qu**, UIUC, United States of America
Co-author(s): Fei Xue
Schedule: *Thursday 14 June, 15:00–15:30*

Abstract: Penalty-based variable selection methods are powerful in selecting relevant covariates and estimating coefficients simultaneously. However, variable selection could fail to be consistent when covariates are highly correlated. The partial correlation approach has been adopted to solve the problem with correlated covariates. Nevertheless, the restrictive range of partial correlation is not effective for capturing signal strength for relevant covariates. In this paper, we propose a new Semi-standard PArtial Covariance (SPAC) which is able to reduce correlation effects from other predictors while incorporating the magnitude of coefficients. The proposed SPAC variable selection facilitates choosing covariates which have direct association with the response variable, via utilizing dependency among covariates. We show that the proposed method with the Lasso penalty (SPAC-Lasso) and the SCAD penalty (SPAC-SCAD) enjoys strong sign consistency in both finite-dimensional and high-dimensional settings under regularity conditions. Simulation studies and the 'HapMap' gene data application show that the proposed method outperforms the traditional Lasso, adaptive Lasso, SCAD, and Peter–Clark-simple (PC-simple) methods for highly correlated predictors.

Title: ***Significance testing in non-sparse high-dimensional linear models***
Speaker: **Yinchu Zhu**, University of Oregon, United States of America
Co-author(s): Jelena Bradic
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: In high-dimensional linear models, the sparsity assumption is typically made, stating that most of the parameters are equal to zero. Under the sparsity assumption, estimation and, recently, inference have been well studied. However, in practice, sparsity assumption is not checkable and more importantly is often violated, with a large number of covariates expected to be associated with the response, indicating that possibly all, rather than just a few, parameters are non-zero. In this article, we propose a new inferential method, named CorrT, which is robust to model misspecification and adaptive to the sparsity assumption. CorrT is shown to have Type I error approaching the nominal level for *any* models and Type II error approaching zero for sparse and many dense models. In fact, CorrT is also shown to be optimal in a variety of frameworks: sparse, non-sparse and hybrid models where sparse and dense signals are mixed. Numerical experiments show a favorable performance of the CorrT test compared to the state-of-the-art methods.

Title: ***Prediction and Inference under Competing Risks in High Dimension - An EHR Demonstration Project for Prostate Cancer***

Speaker: **Ronghui Xu**, University of California, San Diego, United States of America
Schedule: *Thursday 14 June, 16:00–16:30*
Abstract: Our work was motivated by the analysis project using the linked SEER-Medicare database to predict cancer versus non-cancer mortality in men of age 65 years or older who were diagnosed with prostate cancer. We consider existing R package implementations that are computationally feasible for such data sets with up to 100,000 human subjects and over 20,000 claim codes. We carried out simulation studies to compare lasso for cause-specific hazards regression and boosting for Fine-Gray sub-distribution regression models with different approaches to choosing the penalty parameters, with the goal of prediction accuracy of the cumulative incidence rates at 2 and 5 years from baseline. Separately, we also developed methods for constructing confidence intervals of the regression effects of predictors (clinical variables and claim codes) under the Fine-Gray model in high dimensions, using the one-step estimator along the line of van de Geer et al. (2014) and Zhang and Zhang (2014). The extension though, is non-trivial under competing risks and Cox type model formulation. The results of the SEER-Medicare database analysis will be presented.

Invited Session IS-55 • Thursday 14, 14:30-16:30 • Room Amalfi

NONPARAMETRIC MODELS FOR SPATIO-TEMPORAL DATA
Organizer(s): Stefano Castruccio (University of Notre Dame)
Chair: Simone Padoan (Bocconi University)

Title: ***Asymptotic results for cross validation estimation of covariance parameters***
Speaker: **François Bachoc**, Institut de Mathématiques de Toulouse, France
Co-author(s): Agnes Lagnoux, Thi Mong Ngoc Nguyen
Schedule: *Thursday 14 June, 14:30–15:00*
Abstract: We will present some recent asymptotic results on cross validation estimation of covariance parameters for Gaussian process models. We will first address the well-specified case, where the true covariance function does belong to the parametric family of covariance functions used for estimation. In this case, we will provide a consistency and asymptotic normality result for cross validation under fixed-domain asymptotics, in the case of the exponential

covariance function in dimension one. The asymptotic variance depends on the triangular array of observation points and can be equal to that of maximum likelihood in some cases, and up to twice as large in other cases. Second, we will address the misspecified case under increasing-domain asymptotics. We will provide a general result showing that cross validation can be optimal for the integrated square prediction error.

Title: ***Regression Type Models for Extremal Dependence***
Speaker: **Miguel de Carvalho**, University of Edinburgh, United Kingdom
Co-author(s): Linda Mhalla, Valérie Chavez-Demoulin
Schedule: *Thursday 14 June, 15:00–15:30*
Abstract: In this talk, I will discuss a vector generalized additive modelling framework for taking into account the effect of covariates on angular density functions in a multivariate extreme value context. The proposed methods are tailored for settings where the dependence between extreme values may change with time or other covariates. I will devise a maximum penalized log-likelihood estimator, discuss details of the estimation procedure, and on its consistency and asymptotic normality. The empirical analysis reveals relevant dynamics of the dependence between extreme air temperatures in two alpine resorts during the winter season.

Title: ***Spatio-temporal Functional Data Analysis***
Speaker: **Matthew Reimherr**, Penn State University, United States of America
Co-author(s): Piotr Kokoszka, Gregory Bopp, John Ensley
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: In this presentation we discuss several recent developments concerning statistical inference for spatio-temporal functional data. An overview of the structure of the data and models will be presented, followed by methods for testing for independence, change-points, and separability. The methods will be illustrated on data collected across Russian meteorological stations.

Title: ***Predicting missing values in spatio-temporal satellite data***
Speaker: **Reinhard Furrer**, University of Zurich, Switzerland
Co-author(s): Florian Gerber
Schedule: *Thursday 14 June, 16:00–16:30*
Abstract: Remotely sensed data are sparse, which means that data have missing values, for instance due to cloud cover. This is problematic for applications and signal processing algorithms that require complete data sets. To address the sparse data issue, we present a new gap-fill algorithm. The proposed method predicts each missing value

separately based on data points in a spatio-temporal neighborhood around the missing data point. The computational workload can be distributed among several computers, making the method suitable for large datasets. The prediction of the missing values and the estimation of the corresponding prediction uncertainties are based on sorting procedures and quantile regression. The algorithm was applied to MODIS NDVI data from Alaska and tested with realistic cloud cover scenarios featuring up to 50% missing data. Validation against established software showed that the proposed method has a good performance in terms of the root mean squared prediction error. The procedure is implemented and available in the open-source R package gapfill. We demonstrate the software performance with a real data example and show how it can be tailored to specific data.

Invited Session IS-56 • Thursday 14, 14:30-16:30 • Room Procida

NONPARAMETRIC CONDITIONAL MODELING: COPULAS, KERNELS, LOCAL GAUSS, NETWORKS

Organizer(s): Dag Bjarne Tjøstheim (University of Bergen)
Chair: Bård Støve (University of Bergen)

Title: ***Characterizing conditional dependence and testing for conditional independence using the local Gaussian partial correlation***

Speaker: **Håkon Otneim**, NHH Norwegian School of Economics, Norway

Co-author(s): Dag Tjøstheim

Schedule: *Thursday 14 June, 14:30–15:00*

Abstract: Two random variables X and Y are conditionally independent given a set of variables Z if $X|Z$ and $Y|Z$ are independent. This concept is important in many applications, such as regression analysis and when building Bayesian networks, and has received much attention in the literature lately. Just as the correlation coefficient characterizes dependence between jointly Gaussian variables, the partial correlation coefficient characterizes conditional dependence between jointly Gaussian variables, but outside the multivariate normal distribution these statements do not hold. We will therefore define the local Gaussian partial correlation function which in a large class of distributions, given a set of variables Z, is zero on the sample space of (X,Y) if and only if X and Y are conditionally independent given Z. The partial correlation can be estimated using a local likelihood procedure, and the estimate forms the basis of a new and

powerful test for conditional independence. We will look at some theoretical aspects as well as simulation results and applications on real data.

Title: ***Nonparametric estimation of pair-copula constructions with the empirical pair-copula***
Speaker: **Ingrid Hobæk Haff**, University of Oslo, Norway
Co-author(s): Johan Segers
Schedule: *Thursday 14 June, 15:00–15:30*
Abstract: A pair-copula construction is a decomposition of a multivariate copula into a structured system, called regular vine, of bivariate copulae or pair-copulae. The standard practice is to model these pair-copulae parametrically, inducing a model risk, with errors potentially propagating throughout the vine structure. The empirical pair-copula provides a nonparametric alternative, which is conjectured to still achieve the parametric convergence rate. Its main advantage for the user is that it does not require the choice of parametric models for each of the pair-copulae constituting the construction. It can be used as a basis for inference on dependence measures, for selecting an appropriate vine structure, and for testing for conditional independence.

Title: ***Nonparametric vine models with application***
Speaker: **Claudia Czado**, Technical University of Munich , Germany
Co-author(s): Thomas Nagler
Schedule: *Thursday 14 June, 15:30–16:00*
Abstract: Vine copula models based on parametric pair copulas have been shown to provide a very flexible multivariate copula class. However some data sets might require additional flexibility since they exhibit dependence structures in the pair copulas which cannot be captured by parametric pair copulas. Therefore several approaches have been pursued to substitute parametric pair copula families by nonparametric fits. In particular kernel based estimators show better performance than B-splines based methods when the dependence is stronger. We introduce the methods and give applications show casing the advantages over parametric vine models.
References:
Nagler, T., & Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. Journal of Multivariate Analysis, 151, 69-89.
Nagler, T., Schellhase, C., & Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. Dependence Modeling, 5(1), 99-120.

Title: ***Shape Constrained Nonparametric Instrumental Regression***
Speaker: **Jeffrey S. Racine**, McMaster University, Canada
Schedule: *Thursday 14 June, 16:00–16:30*
Abstract: We consider nonparametric kernel estimation of an instrumental regression function φ defined by conditional moment restrictions that stem from a structural econo- met- ric model $E(Y - \varphi(Z)|W) = 0$, and involve endogenous variables Y and Z and instruments W . The function φ is the solution to an ill-posed inverse problem. Our primary focus lies in shape constrained estimation, and we present a simple and robust approach towards shape constrained local nonparametric instrumental regression. The constraints can be imposed on the estimated function $\hat{\varphi}$, its derivatives, or combinations thereof. Our approach facili- tates imposing, say, axioms of consumer/producer theory on an otherwise unrestricted but smooth estimate. Theoretical underpinnings are provided and applications are considered.

Contributed Session CS-24 • Thursday 14, 17:00-18:00 • Room Ravello

NONPARAMETRIC TEST AND MODELLING II

Chair: Carlos Brunet Martins-Filho (University of Colorado at Boulder)

Title: ***A Nonparametric Goodness-Of-Fit Test***
Speaker: **Dimitrios Bagkavos**, University of Crete, Greece
Co-author(s): Prakash Patil, Andrew Wood
Schedule: *Thursday 14 June, 17:00–17:20*
Abstract: We revisit the goodness-of-fit test of Bickel and Rosenblatt (1973) and note the slow rate of its test statistic's convergence to the normal distribution. We then propose an alternative procedure, but still based on smoothing methodology, to test the goodness-of-fit of a given probability model. The test statistic of the proposed test does converge to the normal distribution with a rate faster than that of the statistic considered in Bickel and Rosenblatt (1973) but more importantly the test has the power against a large class of alternatives and can detect alternatives converging to null at a rate $N^{-\beta}$ where $\beta < 1/2$ and N the sample size.

Title: ***Nonparametric estimation of the missing mass in feature allocation models***
Speaker: **Marco Battiston**, Oxford University, United Kingdom
Schedule: *Thursday 14 June, 17:20–17:40*

Abstract: Feature allocation models are generalization of clustering and species sampling models and have recently been used in many fields, like for instance machine learning, ecology and genetics. In some of these applications, interest lies in estimating the expected number of new features to be observed in future observations. This information is useful to do sample size determination and to estimate measures of feature variety in the population. A nonparametric estimator is proposed for the expected number of new features to be observed in the next observation. This estimator resembles the famous Good-Turing estimator for species sampling models. It can be derived as a nonparametric Empirical Bayes estimator, using an Indian Buffet Process as prior and estimating its hyperparameters with consistent frequentist estimators. For the resulting EB estimator, tight confidence intervals are derived and good empirical performances are shown through simulated and real data examples. Moreover, the proposed estimator is shown to achieve the minimax rate under an adequate loss function and when the parameter space is restricted to feature probabilities having tail heavy enough. The proof of this result relies on lower bounding the minimax risk using tools from Bayesian Nonparametrics, which can be of independent interest.

Title: ***Nonparametric estimation of unrestricted distribution functions***

Speaker: **Carlos Brunet Martins-Filho**, Kazakh-British Technical University, Kazakhstan

Co-author(s): Kairat Mynbaev

Schedule: *Thursday 14 June, 17:40–18:00*

Abstract: We consider nonparametric estimation of an extit{ unrestricted } distribution F in that that it may, or may not, be absolutely continuous. First, for x a point of continuity of F, we consider estimators that can be expressed as $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right)$, for a suitable choice of U and bandwidth $h > 0$, and obtain their rates of convergence to $F(x)$. Contrary to the extant literature, we make no restriction on the existence or smoothness of the derivatives of F. The key insight for our result is the use of Lebesgue-Stieltjes integrals. A special case of $\hat{F}_n(x)$, that reproduces the traditional kernel estimator, is given when $U(x) = \int_x^{\infty} K(u)du$ and K is a kernel. Second, for x that is either a point where F has a jump discontinuity, or isolated, we obtain rates of convergence for an estimator $\hat{J}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h}\right)$ for the jump $F(x) - F(x-)$ and suitable choice of W. Once again, no restriction is imposed on F beyond right-continuity. Here, a suitable choice is $W(x) = \int_{\mathbb{R}} e^{ixu} K(u)du$, the Fourier transform of a kernel

K. Our result are of significant practical use as there are numerous examples in Economics, Finance and Biomedicine of distributions that have mass points and singularities [Zinde-Walsh (2008, extit{ Econometric Theory } 24, 696-725)].

Contributed Session CS-25 • Thursday 14, 17:00-18:00 • Room Furore

FUNCTIONAL DATA II

Chair: José R. Berrendero (Universidad Autónoma de Madrid)

Title: ***Independent and conditionally independent counterfactual distributions***

Speaker: **Marcin Wolski**, European Investment Bank, Luxembourg

Schedule: *Thursday 14 June, 17:00–17:20*

Abstract: This study proposes a novel fully nonparametric dependence filtering framework. The filter provides a draw from a counterfactual distribution which is independent, or conditionally independent, from the effects of given covariates. We provide estimation techniques and an inference roadmap for counterfactual distributions, and a numerical exercise confirms that the approach performs well in nonlinear environments. Furthermore, we provide bootstrap validity results for the confidence interval of the estimates. We apply the approach to filter out the sovereign risk spill-overs on corporate cost of borrowing in selected euro area countries.

Title: ***Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes***

Speaker: **Yoav Zemel**, Georg-August-Universität Göttingen, Germany

Co-author(s): Valentina Masarotto, Victor Panaretos

Schedule: *Thursday 14 June, 17:20–17:40*

Abstract: Covariance operators are fundamental in functional data analysis, providing the canonical means to analyse functional variation via the celebrated Karhunen–Loeve expansion. These operators may themselves be subject to variation, for instance in contexts where multiple functional populations are to be compared. Statistical techniques to analyse such variation are intimately linked with the choice of metric on covariance operators, and the intrinsic infinite-dimensionality of these operators. We describe the manifold-like geometry of the space of trace-class infinite-dimensional covariance operators and associated key statistical properties, under the recently proposed infinite-dimensional version of the Procrustes metric.

We identify this space with that of centred Gaussian processes equipped with the Wasserstein metric of optimal transportation. The identification allows us to provide a detailed description of those aspects of this manifold-like geometry that are important in terms of statistical inference; to establish key properties of the Frechet mean of a random sample of covariances; and to define generative models that are canonical for such metrics and link with the problem of registration of warped functional data.

Title: ***Functional logistic regression and reproducing kernel Hilbert spaces***

Speaker: **José R. Berrendero**, Universidad Autónoma de Madrid, Spain

Co-author(s): Beatriz Bueno-Larraz, Antonio Cuevas

Schedule: *Thursday 14 June, 17:40–18:00*

Abstract: We propose a functional logistic regression model to explore the relationship between a dichotomous response variable and a functional predictor. The proposal is based on ideas borrowed from the theory of reproducing kernel Hilbert spaces (RKHS). Similarly to the finite-dimensional case our model holds when the conditional distributions of the predictor given the two possible values of the response are Gaussian with the same covariance structure. Moreover, some particular choices of the slope function lead to point-impact models. We also give conditions (which include Brownian-like predictors) under which the maximum likelihood estimator of the slope function does not exist with probability one and address some possible solutions.

Contributed Session CS-26 • Thursday 14, 17:00-18:00 • Room Positano

SMOOTHING TECHNIQUES II

Chair: Eric Matzner-lober (CEPE ENSAE)

Title: ***Using jittering to deal with discrete variables in nonparametric regression and density estimation***

Speaker: **Thomas Nagler**, Technical University of Munich, Germany

Schedule: *Thursday 14 June, 17:00–17:20*

Abstract: Most nonparametric function estimators can only handle continuous data. One can show that making discrete variables continuous by adding noise is justified under suitable conditions on the noise distribution. This principle is widely applicable, including density and regression function estimation. An in-depth analysis of the

jittering kernel density estimator reveals several appealing properties regarding bias, convergence rates, and efficiency.

Title: ***Nonparametric panel data models with cross-sectional dependence***
Speaker: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-author(s): Juan M. Rodriguez, Peter M. Robinson
Schedule: *Thursday 14 June, 17:20–17:40*
Abstract: The asymptotic distribution for the local linear estimator in nonparametric panel data regression models is established when cross-sectional dependence is accounted for. More efficient regression techniques are proposed. Further, sufficient conditions for the asymptotic normality of the resulting estimators are given and their efficiency gains relative to the standard nonparametric techniques are established. Asymptotically optimal bandwidth choices are justified for these estimates. Feasible optimal bandwidths, and feasible optimal regression estimates, are also asymptotically justified. The proposed estimators are augmented by a Monte Carlo study and they are also illustrated in an empirical analysis about the relationship between public debt, monetary policy and economic growth for eurozone countries.

Title: ***Boosting k-nn type smoothers***
Speaker: **Eric Matzner-lober**, CEPE ENSAE, France
Co-author(s): P. A. Cornillon, N. Hengartner, T. Kerdreux
Schedule: *Thursday 14 June, 17:40–18:00*
Abstract: The aim of the article is to study some transformations of smoothers used in L_2 boosting. Symmetric smoother with eigen values in $[0,1]$ (such as smoothing spline) have nice properties but some classical smoothers don't have these properties and using them in an iterated framework could be problematic. More specifically, we give a systematic method to transform smoothers into possible iterated estimators. We will focus on k-nearest neighbor type (classical, mutual, symmetric) smoothers and we propose an estimator symmetric with eigen values in $[0,1]$ which could be evaluated at any points.

Special Invited Session SIS-07 • Thursday 14, 17:00-18:00 • Auditorium Tafuri

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DEMAND MODELS

by Yuichi Kitamura
Yale University, USA

Speaker: Yuichi Kitamura
Chair: Stefan Sperlich (University of Geneva)

Abstract: This talk discusses a new approach to analyze demand data by leveraging on restrictions implied by models of consumer choice. Though it involves optimization in a very high — easily exceeding one million — dimensions, with an appropriately formulated algorithm the method can be straightforwardly implemented at a reasonable computational cost. The issue of set-identification and associated non-regularity of statistical inference arise, and practical tools are developed to address them. The proposed approach avoids ad hoc assumptions, and in particular, the nature of unobserved components are left unrestricted. Moreover, it enables the researcher to calculate counterfactual predictions and evaluate policy effects. The framework is general, and it is applicable to standard/non-standard choice theory or problems with strategic interactions among multiple agents as well.

Invited Session IS-57 • Friday 15, 08:30-10:30 • Room Vietri

BOOTSTRAP FOR TIME SERIES

Organizer(s): Efstathios Paparoditis & Jens-Peter Kreiss (University of Cyprus & TU Braunschweig)
Chair: Efstathios Paparoditis (University of Cyprus)

Title: ***Bootstrapping factor models with cross sectional dependence***
Speaker: **Silvia Goncalves**, McGill University, Canada
Schedule: *Friday 15 June, 08:30–09:00*

Abstract: We consider bootstrap methods for factor-augmented regressions with cross sectional dependence among idiosyncratic errors. This is important to capture the bias of the OLS estimator derived recently by Gonçalves and Perron (2014). We first show that a common

approach of resampling cross sectional vectors over time is invalid in this context because it induces a zero bias. We then propose the cross-sectional dependent (CSD) bootstrap where bootstrap samples are obtained by taking a random vector and multiplying it by the square root of a consistent estimator of the covariance matrix of the idiosyncratic errors. We show that if the covariance matrix estimator is consistent in the spectral norm, then the CSD bootstrap is consistent, and we verify this condition for the thresholding estimator of Bickel and Levina (2008). Finally, we apply our new bootstrap procedure to forecasting inflation using convenience yields as recently explored by Gospodinov and Ng (2013).
Keywords: factor model, bootstrap, asymptotic bias.

Title: ***Convolved subsampling and the block bootstrap***
Speaker: **Dan Nordman**, Iowa State University, United States of America
Co-author(s): Johannes Tewes, Dimitris Politis
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: The block bootstrap and subsampling are two philosophically different resampling approaches, both aiming to approximate sampling distributions from dependent data by re-using data blocks. We use a structural connection to subsampling in order to characterize the block bootstrap in a new and general manner. In the fundamental case of sample means, the block bootstrap distribution of a sample mean equals the k-fold self-convolution of a subsampling distribution. It turns out that simple conditions are possible for showing that a convolved subsampling estimator produces normal limit that matches the target of bootstrap estimation. From this relationship, the block bootstrap for means can be established more easily, and under much weaker assumptions, than previously considered in many dependence settings. Beyond sample means, convolved subsampling does not necessarily match the block bootstrap, but instead provides an alternative hybrid-type of resampling estimator. Under minimal dependence conditions, results also establish the validity of such convolved subsampling for general statistics having normal limits.

Title: ***Frequency Domain Hybrid Bootstrap for General Stationary Processes***
Speaker: **Marco Meyer**, Helmut Schmidt University Hamburg
Co-author(s): Efstathios Paparoditis, Jens-Peter Kreiss
Schedule: *Friday 15 June, 09:30–10:00*
Abstract: For time series the class of spectral mean estimators (also called integrated periodograms) includes many important statistics such as

sample autocovariances and autocorrelations, as well as smoothed spectral density estimators, among other things. Existing methods for bootstrapping these estimators have a very limited range since they are not able to capture the dependence structure of periodogram ordinates at different frequencies for finite sample sizes. Essentially, these procedures cover the case of univariate, linear time series with independent innovations, and some even require the time series to be Gaussian. We propose a new, hybrid bootstrap method in the frequency domain which is consistent for a much wider range of stationary processes. It uses existing methods to replicate the dominant part of the distribution of interest, and employs a new concept of convolved subsampling to correct for those features of the distribution of interest that cannot be mimicked by classical procedures. We show consistency for this hybrid procedure for a general class of time series, ranging clearly beyond linear processes, and for general spectral mean statistics. This yields, as one example, a method to consistently bootstrap sample autocovariances in non-linear processes. The performance of this procedure is illustrated via simulations.

Title: ***Testing Equality of Spectral Density Operators for Functional Linear Processes***

Speaker: **Theofanis Sapatinas**, University of Cyprus, Cyprus

Co-author(s): Anne Leucht, Efstathios Paparoditis

Schedule: *Friday 15 June, 10:00–10:30*

Abstract: The problem of testing equality of the entire second order structure of two independent functional linear processes is considered. A fully functional L^2-type test is developed which evaluates, over all frequencies, the Hilbert-Schmidt distance between the estimated spectral density operators of the two processes. The asymptotic behavior of the test statistic is investigated and its limiting distribution under the null hypothesis is derived. Furthermore, a novel frequency domain bootstrap method is developed which approximates more accurately the distribution of the test statistic under the null than the large sample Gaussian approximation obtained. Asymptotic validity of the bootstrap procedure is established and consistency of the bootstrap-based test under the alternative is proved. Numerical simulations show that, even for small samples, the bootstrap-based test has very good size and power behavior. An application to meteorological functional time series is also presented.

DIRECTIONAL DATA

Organizer(s): Marco Di Marzio (Università degli studi di Chieti-Pescara)
Chair: Marco Di Marzio (Università degli studi di Chieti-Pescara)

Title: ***A bandwidth-free kernel estimator of the integral of a squared density***
Speaker: **Jose E. Chacon**, Universidad de Extremadura, Spain
Schedule: *Friday 15 June, 08:30–09:00*
Abstract: Given the density f of a circular random variable Θ, the goal is to estimate the integral $\psi_0 = \int_0^{2\pi} f(\theta)^2 d\theta$. This problem has attracted much interest for linear data, and applications can be found in the introductory sections of Aldershof (1991) or, more recently, Ginè and Nickl (2008). For circular variables, however, the problem has received far less attention. The asymptotic properties of a natural kernel-type estimator can be deduced from a general result contained in Di Marzio, Panzera and Taylor (2011), which shows that the optimal bandwidth for estimating $\psi_r = \int_0^{2\pi} f^{(r)}(\theta) f(\theta) d\theta$ depends on ψ_{r+2}, so that the problem becomes somehow cyclic. In the linear case, the usual solution consists of using a multi-stage estimation procedure, starting with an initial approximation based on some reference distribution. Unfortunately, for circular data, the common choice of the von Mises distribution as the reference distribution is known to entail serious difficulties, for example in the presence of antipodal modes. Here we present a new estimator for ψ_0 which, despite being based on kernel smoothing, does not require a bandwidth choice or reference distributions. Even so, the estimator is shown to be asymptotically efficient and very competitive in practice.

Title: ***On optimal tests for rotational symmetry against new classes of hyperspherical distributions***
Speaker: **Eduardo García-Portugués**, Carlos III University of Madrid, Spain
Co-author(s): Davy Paindaveine, Thomas Verdebout
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: Motivated by the central role played by rotationally symmetric distributions in directional statistics, we consider the problem of testing rotational symmetry on the hypersphere. We adopt a semiparametric approach and tackle the situations where the location of the symmetry axis is either specified or unspecified. For each problem, we define two tests and study their asymptotic properties

under very mild conditions. We introduce two new classes of directional distributions that extend the rotationally symmetric class and are of independent interest. We prove that each test is locally asymptotically maximin, in the Le Cam sense, for one kind of the alternatives given by the new classes of distributions, both for specified and unspecified symmetry axis. The tests, aimed to detect location-like and scatter-like alternatives, are combined into a convenient hybrid test that is consistent against both alternatives.

Title: ***Directional data depths with applications***
Speaker: **Giovanni C. Porzio**, University of Cassino and Southern Lazio, Italy
Schedule: *Friday 15 June, 09:30–10:00*
Abstract: Depth functions allow an inner-outer ordering of data in multivariate spaces with respect to a sample or a distribution. Depth functions have been defined for directional data as well, with many functions available within the literature. Along with the first introduced angular Tukey's, angular simplicial, and arc distance depths, more computationally feasible functions have been more recently defined: the angular Mahalanobis depth, based on the idea of a projection ordering, and the cosine and the chord distance depths, with the latters belonging to a class based on directional distances. Within such a setting, this work presents a systematic review of the directional depth functions available, and their potential applications in directional statistics. Particularly, their use to obtain directional location estimators and to define a directional classifier is investigated. A supervised classification tool is obtained adopting the maxdepth and the DD-plot classifier approach.

Title: ***Kernel methods for errors-in-variables problems in the circular domain***
Speaker: **Stefania Fensore**, Università degli studi di Chieti-Pescara, Italy
Co-author(s): Marco Di Marzio, Agnese Panzera
Schedule: *Friday 15 June, 10:00–10:30*
Abstract: We consider the problem of nonparametrically estimating a circular density or a regression with a circular predictor from data contaminated by angular errors. Specifically, by following the unbiased scores approach, we obtain nonparametric estimators for the functions at hand starting from kernel-based estimators in the error-free case. By formulating the unbiased scores equations in the Fourier domain, we find that the resulting estimators, despite sharing the structure of the corresponding ones for the error-free case, are characterized by weights which are reminiscent of deconvolution kernels. Here, differently from the Euclidean setting,

discrete Fourier coefficients are involved rather than characteristic functions. We derive asymptotic properties of the proposed estimators, provide some simulation results, and also consider some possible generalizations and extensions.

Invited Session IS-59 • Friday 15, 08:30-10:30 • Room Furore

STATISTICAL PROBLEMS INVOLVING BIG DATA

Organizer(s): George Michailidis (University of Florida)
Chair: George Michailidis (University of Florida)

Title: ***Locating targets via wireless sensor networks***
Speaker: **Rohit Patra**, University of Florida, United States of America
Co-author(s): Moulinath Banerjee, George Michailidis
Schedule: *Friday 15 June, 08:30–09:00*
Abstract: Wireless sensor networks (WSNs) serve as key technological infrastructure for monitoring diverse systems across space and time. Examples of their widespread applications include: precision agriculture, surveillance, animal behavior, drone tracking, and emergent disaster response and recovery. A WSN consists of hundreds or thousands of identical sensors at fixed locations where each individual sensor observes the surrounding at fixed time intervals. In this work we estimate the location of a (signal emitting) target under the assumption that magnitude of signal detected at the sensor is a strictly decreasing function of the distance between the sensor and the signal emitting target. We propose an automated root-n-consistent estimator of the location the target under under only the monotonicity assumption. Our estimator is tuning parameter free. We show that our estimator has a Gaussian limit distribution and construct asymptotic confidence region for the location target. This is a joint work with George Michailidis and Moulinath Banerjee.

Title: ***Accurate Change Point Detection in Large Random Graphs via Sampling***
Speaker: **Moulinath Banerjee**, University of Michigan, USA
Co-author(s): Mingyuan Gao, George Michailidis
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: Consider a large random graph with several million edges where the status of each edge evolves in time (present/absent) with its own probability parameter. At an unknown time point, the probability parameter changes for a certain unknown proportion of edges, and

the graph keeps evolving as before. With such large graphs, the change-point in time can often be identified exactly with probability going to 1, but a statistical analysis of the behavior of the entire graph over the observed time domain (say, via least squares) can be computationally expensive. Under certain mild conditions, we develop an inference scheme that allows accurate detection of the change-point via randomly sampling an appropriate number of edges from the graph (depending on its intrinsic parameters which are estimated); in particular, we provide an analytical formula that guarantees a pre-specified level of precision to the inference. We present extensive simulation results to illustrate the effectiveness of our approach. As the number of sampled edges is of a much smaller order compared to the total number of possible edges, this provides comprehensive computational gains without sacrificing statistical accuracy. We also present a competing doubling algorithm that shows impressive performance.

Title: ***Scalable kernel-based variable selection with sparsistency***

Speaker: **Junhui Wang**, City University of Hong Kong, Hong Kong S.A.R. (China)

Schedule: *Friday 15 June, 09:30–10:00*

Abstract: Variable selection is central to sparse modeling, and many methods have been proposed under various model assumptions. In this talk, we will present a scalable framework for model-free variable selection in reproducing kernel Hilbert space (RKHS) without specifying any restrictive model. As opposed to most existing model-free variable selection methods requiring fixed dimension, the proposed method allows dimension p to diverge with sample size n. The proposed method is motivated from the classical hard-threshold variable selection for linear models, but allows for general variable effects. It does not require specification of the underlying model for the response, which is appealing in sparse modeling with a large number of variables. The proposed method can also be adapted to various scenarios with specific model assumptions, including linear models, quadratic models, as well as additive models. The asymptotic estimation and variable selection consistencies of the proposed method are established in all the scenarios.

ADVANCES IN FINANCIAL MODELLING

Organizer(s): Genaro Sucarrat (BI Norwegian Business School)
Chair: Genaro Sucarrat (BI Norwegian Business School)

Title: ***Asymptotics of Cholesky GARCH Models and Time-Varying Conditional Betas***
Speaker: **Christian Francq**, CREST-ENSAE, France
Co-author(s): Serge Darolles, Sébastien Laurent
Schedule: *Friday 15 June, 08:30–09:00*
Abstract: This paper proposes a new model with time-varying slope coefficients. Our model, called CHAR, is a Cholesky-GARCH model, based on the Cholesky decomposition of the conditional variance matrix introduced by Pourahmadi (1999) in the context of longitudinal data. We derive stationarity and invertibility conditions and prove consistency and asymptotic normality of the Full and equation-by-equation QML estimators of this model. We then show that this class of models is useful to estimate conditional betas and compare it to the approach proposed by Engle (2016). Finally, we use real data in a portfolio and risk management exercise. We find that the CHAR model outperforms a model with constant betas as well as the dynamic conditional beta model of Engle (2016).

Title: ***Heterogeneous Component MEM models for forecasting trading volumes***
Speaker: **Giuseppe Storti**, Università degli studi di Salerno
Co-author(s): Antonio Naimoli
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: We propose a novel approach for modelling and forecasting high-frequency trading volumes, revisiting the Component Multiplicative Error Model of Brownlees et al. (2011) by a more flexible specification of the long-run component which is based on a Heterogeneous MIDAS polynomial structure. This uses an additive cascade of MIDAS polynomial filters moving at different frequencies in order to reproduce the changing long-run level and the persistent autocorrelation structure of high frequency trading volumes. The merits of the proposed approach are illustrated by means of an application to three stocks traded on the XETRA market characterised by different degrees of liquidity.

Title: ***Virtual Historical Simulation of the conditional VaR of the returns***

of a portfolio

Speaker: **Jean-Michel Zakoian**, CREST, France

Schedule: *Friday 15 June, 09:30–10:00*

Abstract: In order to estimate the conditional risk of the portfolio's return, two strategies can be advocated. A multivariate strategy requires a dynamic model for the vector of risk factors, but it faces the dimensionality curse. A univariate approach based on a dynamic model for the portfolio's return seems more attractive. However, when the portfolio's return is a time varying combination of the individual returns, we show that this "naive" approach is actually invalid. Instead, a reasonable approach consists in reconstituting a "virtual portfolio", whose returns are built using the current composition of the portfolio. We call this approach Virtual Historical Simulation (VHS). This paper establishes the asymptotic properties of the VHS method. Numerical illustrations on simulated and real data are provided.

Title: ***Volatility Estimation when Observations Are Missing***

Speaker: **Genaro Sucarrat**, BI Norwegian Business School, Norway

Co-author(s): Natalia Bahamonde, Hamdi Raissi

Schedule: *Friday 15 June, 10:00–10:30*

Abstract: In empirical practice observations are often missing. This invalidates standard estimation methods of Generalised Autoregressive Conditional Heteroscedasticity (GARCH) models because of a repeated invertibility issue induced at each missing location. To sidestep this problem we propose a log-ARCH model without GARCH terms, but with stochastic conditioning covariates (e.g. volatility proxies) to compensate for the missing GARCH terms. Apart from the omitted GARCH terms, however, the model is very general and flexible: It is asymmetric, multivariate, allows for (unknown) Dynamic Conditional Correlations (DCCs), and non-negativity constraints are not needed on the parameters nor on the covariates. We derive a least squares equation-by-equation estimator of the model, and prove its Consistency and Asymptotic Normality (CAN) when the missing data process is stationary, unknown and not necessarily independent of the log-ARCH process itself. Our results are illustrated in a simulation study, and in an empirical application.

APPROXIMATION, LEARNING AND INFERENCE WITH POSITIVE-DEFINITE KERNELS

Organizer(s): Bharath K. Sriperumbudur (The Pennsylvania State University)

Chair: Bharath K. Sriperumbudur (The Pennsylvania State University)

Title: ***Approximation analysis of deep CNNs and distributed learning***

Speaker: **Ding-Xuan Zhou**, City University of Hong Kong, Hong Kong S.A.R. (China)

Schedule: *Friday 15 June, 08:30–09:00*

Abstract: In this talk we present some approximation analysis for deep convolutional neural networks (CNNs) and distributed learning algorithms. We show that deep CNNs with the rectified linear unit activation function modified in a distributed setting (without any fully-connected layers) are universal as the depth of the deep network tends to infinity. We demonstrate error analysis for some distributed learning algorithms presented in reproducing kernel Hilbert spaces. Our approach is based on machine learning and approximation theory.

Title: ***Multitask Learning for Contextual Bandits***

Speaker: **Clayton Scott**, University of Michigan, United States of America

Schedule: *Friday 15 June, 09:00–09:30*

Abstract: Contextual bandits are a form of multi-armed bandit in which the agent has access to predictive side information (known as the context) for each arm at each time step, and have been used to model personalized news recommendation, ad placement, and other applications. In this work, we propose a kernel-based multi-task learning framework for contextual bandit problems. Like multi-task learning in the batch setting, the goal is to leverage similarities in contexts for different arms so as to improve the agent's ability to predict rewards from contexts. We propose an upper confidence bound-based multi-task learning algorithm for contextual bandits, establish a corresponding regret bound, and interpret this bound to quantify the advantages of learning in the presence of high task (arm) similarity. This is joint work with Aniket Deshmukh and Urun Dogan.

Title: ***Strictly proper kernel scores and characteristic kernels on compact spaces***

Speaker: **Ingo Steinwart**, University of Stuttgart, Germany

Co-author(s): Johanna Ziegel
Schedule: *Friday 15 June, 09:30–10:00*
Abstract: Strictly proper kernel scores are well-known tool in probabilistic forecasting, while characteristic kernels have been extensively investigated in the machine learning literature. We first show that both notions coincide, so that insights from one part of the literature can be used in the other. We then show that the metric induced by a characteristic kernel cannot reliably distinguish between distributions that are far apart in the total variation norm as soon as the underlying space of measures is infinite dimensional. In addition, we provide a characterization of characteristic kernels in terms of eigenvalues and -functions and apply this characterization to the case of continuous kernels on (locally) compact spaces. In the compact case we further show that characteristic kernels exist if and only if the space is metrizable. As special cases of our general theory we investigate translation-invariant kernels on compact Abelian groups and isotropic kernels on spheres.

Invited Session IS-62 • Friday 15, 08:30-10:30 • Room Amalfi

NONPARAMETRIC VARIABLE SELECTION

Organizer(s): Adriano Zanin Zambom (California State University Northridge)
Chair: Adriano Zanin Zambom (California State University Northridge)

Title: ***Covariate selection and dimension reduction in causal inference***
Speaker: **Xavier de Luna**, Umea University, Sweden
Schedule: *Friday 15 June, 08:30–09:00*
Abstract: In order to draw inference on the causal effect of a treatment, observational studies must be designed with care. A commonly used nonparametric design to identify a causal parameter (treatment effect) of interest is to match or control for a large set of covariates which are believed to potentially explain the treatment assignment mechanism and the outcome on which the effect is to be evaluated. However, there is a non-trivial bias-variance trade-off when a choice on which covariates to use has to be made. This trade-off is not trivial because it depends on different aspects of the study design: the population and parameter of interest, the identification strategy, the assumptions made on different nuisance functions, as well as on the estimators used. Arguably, covariate selection and dimension reduction is thus

more challenging in causal inference than in the well studied case of regression modelling for prediction. We will formalize the problem of covariate selection for causal inference. This formalization is essential in order to be able to the discuss data-driven covariate selection methods, and dimension reduction, for the optimal design of observational studies.

Title: ***Nonparametric variable selection with Hamming loss***
Speaker: **Cristina Butucea**, CREST, ENSAE, Université Paris Saclay, France
Co-author(s): Natalia Stepanova, Alexandre Tsybakov
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: We derive non-asymptotic bounds for the minimax risk of variable selection under expected Hamming loss in the Gaussian mean model in Rˆd for classes of s-sparse vectors separated from 0 by a constant a>0. In some cases, we get exact expressions for the non-asymptotic minimax risk as a function of d,s,a and find explicitly the minimax selectors. These results are extended to dependent or non-Gaussian observations and to the problem of crowdsourcing. Analogous conclusions are obtained for the probability of wrong recovery of the sparsity pattern. As corollaries, we derive necessary and sufficient conditions for such asymptotic properties as almost full recovery and exact recovery. Moreover, we propose data-driven selectors that provide almost full and exact recovery adaptively to the parameters of the classes.

Title: ***Strategies for Nonparameteric Variable Selection***
Speaker: **Leonard Allen Stefanski**, NC State University, United States of America
Schedule: *Friday 15 June, 09:30–10:00*
Abstract: The talk will discuss two general strategies for nonparametric variable selection. In the first a nonparametric regression method is combined with a strategy for defining variable importance via measurement error attenuation factors. The second general strategy is based on the idea of tuning selection and shrinkage parameters via the addition of pseudo variables to the data set and tracking their inclusion frequencies. The latter approach derives an estimator of the false selection rate for each model along a solution path and can be used to estimate a model with a pre-specified false selection rate approximately.

Title: ***Hypothesis Testing Sure Independence Screening for Nonparametric Regression***
Speaker: **Adriano Zanin Zambom**, California State University Northridge,

United States of America

Schedule: *Friday 15 June, 10:00–10:30*

Abstract: In this paper we develop a sure independence screening method based on hypothesis testing (HT-SIS) in a general nonparametric regression model. The ranking utility is based on a powerful test statistic for the hypothesis of predictive significance of each available covariate. The sure screening property of HT-SIS is established, demonstrating that all active predictors will be retained with high probability as the sample size increases. The threshold parameter is chosen in a theoretically justified manner based on the desired false positive selection rate. Simulation results suggest that the proposed method performs competitively against procedures found in the literature of screening for several models, and outperforms them in some scenarios. A real dataset of microarray gene expressions is analyzed.

Invited Session IS-63 • Friday 15, 08:30-10:30 • Room Procida

RECENT ADVANCES IN FUNCTIONAL TIME SERIES

Organizer(s): Matthew Logan Reimherr (Penn State University)

Chair: Matthew Reimherr (Penn State University)

Title: ***Inference for the autocovariance of a functional time series under conditional heteroscedasticity***

Speaker: **Greg Rice**, University of Waterloo, Canada

Co-author(s): Piotr Kokoszka, Hanlin Shang

Schedule: *Friday 15 June, 08:30–09:00*

Abstract: Most methods for analyzing functional time series rely at some level on the estimating lagged autocovariance operators or surfaces. We develop methods for performing inference for the lagged autocovariance operators of stationary functional time series that are valid under general conditional heteroscedasticity conditions. These include a portmanteau test to assess the cumulative significance of empirical autocovariance operators up to a user selected maximum lag, as well as methods for obtaining confidence bands for a functional version of the autocorrelation that are useful in model selection/validation. We analyze the efficacy of these methods through a simulation study, and apply them to functional time series derived from asset price data of several representative assets. In this application, we found that strong white noise tests often suggest that such series exhibit significant autocorrelation, whereas our tests,

which account for functional conditional heteroscedasticity, show that these data are in fact uncorrelated in a function space.

Title: ***On Sieve Bootstrap Prediction Intervals for Functional Time Series***
Speaker: **Han Lin Shang**, Australian National University, Australia
Co-author(s): Efstathios Paparoditis
Schedule: *Friday 15 June, 09:00–09:30*
Abstract: A sieve bootstrap procedure for constructing prediction intervals of a functional time series is proposed. The bootstrap procedure exploits a general vector autoregressive representation of the time series of Fourier coefficients appearing in the Karhunen-Loeve expansion of functional process. The bootstrap procedure is developed to generate pseudo functional time series that adequately mimic the dependence structure of the functional time series. The method uses a finite representation of functional principal components to capture the characteristics of the infinite dimensional process and a finite order vector autoregressive process to imitate the temporal dependence structure of the corresponding vector time series of Fourier coefficients. With the pseudo functional time series, we compute the prediction interval via a quantile approach. Through a set of simulation studies and two real-world data sets, we demonstrate good finite-sample performance of the new bootstrap method proposed.

Title: ***A general white noise test based on kernel lag-window estimates of the spectral density operator***
Speaker: **Vaidotas Characiejus**, Université libre de Bruxelles, Belgium
Co-author(s): Gregory Rice
Schedule: *Friday 15 June, 09:30–10:00*
Abstract: We propose a general white noise test for functional time series based on estimating a distance between the spectral density operator of a weakly stationary time series and the constant spectral density operator of an uncorrelated time series. The estimator that we propose is based on a kernel lag-window type estimator of the spectral density operator. When the observed time series is a strong white noise in a real separable Hilbert space, we show that the asymptotic distribution of the test statistic is standard normal, and we further show that the test statistic diverges for general serially correlated time series. These results recover as a special case those of Hong (1996) in the setting of scalar time series. In order to implement the test, we propose and study a number of kernel and bandwidth choices, including a new data adaptive bandwidth for such estimators. A simulation study demonstrated that the proposed method has good size and improved

power when compared to other methods available in the literature, while also offering a light computational burden.

Title: ***Testing normality for functional time series***
Speaker: **Siegfried Hörmann**, Graz University of Technology, Austria
Schedule: *Friday 15 June, 10:00–10:30*
Abstract: We develop tests of normality for time series of functions. The tests are related to the commonly used Jarque–Bera test. The assumption of normality has played an important role in many methodological and theoretical developments in the field of functional data analysis, yet, no inferential procedures to verify it have been proposed so far, even for iid functions. We propose two approaches which handle two paramount challenges: 1) the unknown temporal dependence structure and 2) the estimation of the optimal finite dimensional projection space. We evaluate the tests via simulations and establish their large sample validity under general conditions. We obtain useful insights by applying them to pollution and intraday price curves. While the pollution curves can be treaded as normal, the normality of high frequency price curves is rejected.

Peter Hall Session PL-04 • Friday 15, 11:00-12:00 • Auditorium Tafuri

SPURIOUS CORRELATION AND ITS IMPACT ON ESTIMATION OF ULTRAHIGH DIMENSIONAL NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION MODELS

by Runze Li
Penn State University, USA

Chair: Aurore Delaigle (University of Melbourne)

Abstract: This talk starts with the phenomenon of spurious correlation in ultrahigh dimensional nonparametric models, and an illustration of its impact on error variance estimation of ultrahigh dimensional additive models. Specifically, the traditional mean squared errors significantly underestimates the error variance. I further introduce an accurate estimate for error variance in ultrahigh dimensional sparse additive model by effectively integrating sure independence screening and refitted cross-validation techniques. I will discuss the spurious correlation in semiparametric regression and its impact on the baseline function estimation in ultrahigh dimensional partial linear models.

FUNCTIONAL DATA III

Chair: Jean-Baptiste Aubin (INSA-Lyon)

Title: ***Recursive feature selection with functional data***
Speaker: **José Luis Torrecilla**, Universidad Autónoma de Madrid, Spain
Co-author(s): Alberto Suárez
Schedule: *Friday 15 June, 12:00–12:20*
Abstract: Classifying functional data entails some difficulties related to their high (infinite) dimension, so some kind of dimensionality reduction is often needed. Variable selection techniques are becoming popular in the functional data analysis literature since they lead to an interpretable reduction with good accuracy rates in a subsequent classification. In this context (supervised functional classification), the maxima hunting method (MH) performs variable selection by identifying the maxima of a dependence function between the predictive functional variable and the class label. MH presents a good performance and some valuable properties, however, the relevance of a variable is assessed individually. Here we present a recursive extension of MH which solves this drawback by subtracting the expectation of the process conditioned to the selected maximum. The improvement is illustrated and assessed with both simulations and real data.

Title: ***Functional central limit theorems for the Nelson–Aalen and Kaplan–Meier estimators for dependent stationary data***
Speaker: **Dragi Anevski**, Lund University, Sweden
Schedule: *Friday 15 June, 12:20–12:40*
Abstract: We derive process limit distribution results for the Nelson–Aalen estimator of a hazard function and for the Kaplan–Meier estimator of a distribution function, under different dependence assumptions. The data are assumed to be right censored observations of a stationary time series. We treat weakly dependent as well as long range dependent data, and allow for qualitative differences in the dependence for the censoring times versus the time of interest.
Reference: Anevski, D. "Functional central limit theorems for the Nelson–Aalen and Kaplan–Meier estimators for dependent stationary data", Statistics and Probability Letters, 2017, Vol 124, pp 83-91

Title: ***Local dimensional reduction using small ball probability factorization***

Speaker: **Jean-Baptiste Aubin**, INSA-Lyon, France
Co-author(s): Enea Bongiorno
Schedule: *Friday 15 June, 12:40–13:00*
Abstract: The small–ball probability of a Hilbert valued process is considered. Recent works on its factorization put the lights on a factor that is used to study the local dimension of a Hilbert valued process. This work in progress studies the properties of this factor and introduces an estimator. It turns out that such estimator is consistent and asymptotically normal distributed. Features of such estimator provide insight on the local dimensionality of the process coherently with recent independent works.

Contributed Session CS-28 • Friday 15, 12:00-13:00 • Room Furore

DIMENSION REDUCTION TECHNIQUES

Chair: Anneleen Verhasselt (Hasselt University)

Title: ***Central Quantile Subspace***
Speaker: **Eliana Christou**, University of North Carolina at Charlotte, United States of America
Schedule: *Friday 15 June, 12:00–12:20*
Abstract: Existing dimension reduction techniques focus on the conditional distribution of the response given the covariates, where specific interest focuses on statistical functionals of the distribution, such as the conditional mean, conditional variance and conditional quantile. We introduce a new method for inferring about the conditional quantile of the response given the covariates and we introduce the notion of the Central Quantile Subspace (CQS). First, we focus on cases where the τ-th conditional quantile, for $\tau \in (0,1)$, depends on the predictor X through a single linear combination $B_\tau' X$ and we show that we can estimate B_τ consistently up to a multiplicative scalar, even though the estimate might be based on a misspecified link function. Second, we extend the result to τ-th conditional quantiles that depend on X through a d_τ-dimensional linear combination $B_\tau' X$, where B_τ is a $p \times d_\tau$ matrix, $d_\tau > 1$, and propose an iterative procedure to produce more vectors in the τ-th CQS, which are shown to be root n consistent. Third, we extend our proposed methodology by considering any statistical functional of the conditional distribution and estimate the fewest linear combinations of X that contain all the information on that functional.

Title: ***Feature Selection in L0 Norm: A Viable Approach***
Speaker: **Ana Maria Kenney**, Pennsylvania State University
Co-author(s): Francesca Chiaromonte, Giovanni Felici
Schedule: *Friday 15 June, 12:20–12:40*
Abstract: A new approach to feature selection, which is made possible by improved Mixed Integer Optimization algorithms, is gaining increasing attention in the statistical community. Despite the appeal of this approach, the computational viability on problems of both realistic size (number of features) and complexity (e.g. patterns in the signals, linear associations among features – and in particular across active and non-active ones) is under discussion. We believe that, as proposed, the new approach fails to exploit several facets which could further reduce computational burden and improve performance. This is critical, especially in light of the fact that this approach, like most feature selection techniques, requires data-driven selection of a core tuning parameter. Working within the framework of Mixed Integer Optimization, we put forth simple and effective proposals for improvement, which render tuning computationally viable. Through a carefully designed simulation study and a real data application, we provide successful comparisons with established methods, highlighting pros, cons and avenues for overcoming present limitations.

Title: ***Variable selection in quantile varying coefficient models with heteroscedastic error***
Speaker: **Anneleen Verhasselt**, Hasselt University, Belgium
Co-author(s): Mohammed A. Ibrahim
Schedule: *Friday 15 June, 12:40–13:00*
Abstract: Quantile regression is a great tool to get a thorough view of the relationship between (the distribution of) a response and covariates. We consider a location-scale quantile varying coefficient model with heteroscedastic error to model longitudinal data. In a longitudinal data setting, it is intuitive to allow the coefficients in the varying coefficient model to vary over time. The functional coefficients are estimated with penalized B-splines.

As we allow for heteroscedasticity, the covariates can influence various quantiles of the response differently. Therefore, the problem of variable selection in quantile regression is more challenging. We consider grouped Lasso and nonnegative garrote to perform variable selection in the location as well as the scale. When the problem is high-dimensional a two-stage approach, with a first screening stage (independence screening) is used, before applying grouped Lasso or

nonnegative garrote in the second stage.

Contributed Session CS-29 • Friday 15, 12:00-13:00 • Room Positano

BOOTSTRAP METHODS

Chair: Mihai Giurcanu (University of Chicago)

Title: ***Bootstrapping characteristic functions under local stationarity***
Speaker: **Carina Beering**, Technische Universität Braunschweig, Germany
Co-author(s): Carsten Jentsch, Anne Leucht, Marco Meyer
Schedule: *Friday 15 June, 12:00–12:20*
Abstract: We propose a kernel-type estimator for the local characteristic function of locally stationary processes introduced by Dahlhaus in 1997. Under weak moment conditions, we provide a functional central limit theorem for the local empirical characteristic function process, which we are able to generalize further. Since in some cases asymptotic confidence intervals cannot be computed due to unknown parameters, we use the block bootstrap proposed in Dowla et al. in 2013 to generate a bootstrap estimator for the local CF. Subsequently, we show consistency for this method. Finally, we illustrate the finite sample behaviour of the procedure in a small simulation study for time-varying α-stable distributions for $\alpha \in (1,2)$.
References:
Dahlhaus, R. (1997): Fitting time series models to nonstationary processes. *The Annals of Statistics* 25, 1–37
Dowla, A. et al. (2013). Local block bootstrap inference for trending time series. *Metrika* 76, 733–764

Title: ***Smoothed bootstrap bandwidth selection for nonparametric hazard rate estimation***
Speaker: **Inés Barbeito Cal**, University of A Coruña, Spain
Co-author(s): Ricardo Cao Abad
Schedule: *Friday 15 June, 12:20–12:40*
Abstract: A smoothed bootstrap method is presented for the purpose of bandwidth selection in nonparametric hazard rate estimation for iid data. In this context, two new bootstrap bandwidth selectors are established based on the exact expression of the bootstrap version of the mean integrated squared error of some approximations of the kernel hazard rate estimator. This is very useful since Monte Carlo approximation is no longer needed for the implementation of the two bootstrap selectors. A simulation study is carried out in order to

show the empirical performance of the two new bootstrap smoothing parameter selectors and to compare them with other already existing methods. The methods are illustrated by applying them to a diabetes data set.

Title: ***Bootstrapping LASSO-Type Estimators in Regression Models***
Speaker: **Mihai Giurcanu**, University of Chicago, United States of America
Schedule: *Friday 15 June, 12:40–13:00*
Abstract: We study the consistency of the standard (non-parametric) bootstrap, the m-out-of-n bootstrap, and the oracle bootstrap distributions of some popular LASSO-type estimators in regression models with random predictors. These estimators have an oracle property and are often used in estimation of sparse regression models. A local asymptotic analysis further reveals the behaviour of these estimators and of their bootstrap distributions when some regression coefficients approach zero at various rates. In a simulation study, we assess the finite sample properties of the estimators and of their bootstrap distributions for various sample sizes and model parameters. A statistical analysis of a prostate cancer data set shows an application of LASSO-type inference and bootstrap methods in practice.

Contributed Session CS-30 • Friday 15, 12:00-13:00 • Room Procida

NONPARAMETRIC INFERENCE AND ESTIMATION III
Chair: Ester Mariucci (otto von Guericke universität Magdeburg)

Title: ***Simultaneous inference for curve estimation in time-varying models***
Speaker: **Stefan Richter**, Braunschweig University of Technology, Germany
Co-author(s): Sayar Karmakar, Wei Biao Wu
Schedule: *Friday 15 June, 12:00–12:20*
Abstract: A general class of time-varying regression models which cover general linear models as well as time series models is considered. We estimate the regression coefficients by using local linear M-estimation. For these estimators, Bahadur representations are obtained and are used to construct simultaneous confidence bands. For practical implementation, we propose a bootstrap based method to circumvent the slow logarithmic convergence of the theoretical simultaneous bands. Our results substantially generalize and unify the treatments for several time-varying regression and auto-regression models. The performance for tvARCH and tvGARCH models is

studied in simulations and a few real-life applications of our study are presented through analysis of some financial datasets.

Title: ***EM algorithm for non-parametric mixed-effects models. An application to INVALSI data for unsupervised classification of schools.***
Speaker: **Chiara Masci**, Politecnico di Milano, Italy
Co-author(s): Anna Maria Paganoni, Francesca Ieva
Schedule: *Friday 15 June, 12:20–12:40*
Abstract: We propose an EM algorithm for non-parametric mixed-effects models (NPEM algorithm) and apply it to the National Institute for the Educational Evaluation of Instruction and Training (INVALSI) data of 2013/2014 as a tool for unsupervised clustering of Italian schools. The NPEM algorithm for modelling hierarchical data relaxes the Gaussian assumption of the random effects. In particular, it introduces two main novelties: first it allows the covariates to be group specific; second, it assumes the random effects to be distributed according to a discrete distribution with an (a priori) unknown number of support points. In doing so, it induces an automatic clustering of the grouping factor at higher level of hierarchy, enabling the identification of latent clusters of groups. In the application to INVALSI data, the aim is to identify clusters of schools that differ in their effect on student achievements. The clustering may then be exploited through the use of school level features and geographical factors.

Title: ***Wasserstein and total variation distance for Lévy processes***
Speaker: **Ester Mariucci**, Otto von Guericke Universität Magdeburg, Germany
Co-author(s): Markus Reiß
Schedule: *Friday 15 June, 12:40–13:00*
Abstract: Upper bounds are presented for the Wasserstein distance of order p between the marginals of Lévy processes, including Gaussian approximations for jumps of infinite activity. Using the convolution structure, upper bounds are further derived for the total variation distance between the marginals of Lévy processes. Connections to other metrics like Zolotarev and Toscani-Fourier distances are established. The theory is illustrated by concrete examples and an application to statistical lower bounds.

Special Invited Session SIS-08 • Friday 15, 12:00-13:00 • Auditorium Tafuri

BERNSTEIN-VON MISES TYPE RESULTS FOR MISSPECIFIED NONREGULAR ILL-POSED MODELS

by Natalia Bochkina[(1)] **& Peter J. Green**[(2)]
[(1)]University of Edinburgh, UK; [(2)]University of Technology Sydney, Australia; University of Bristol, UK

Speaker: Natalia Bochkina
Chair: Marianna Pensky (University of Central Florida)

Abstract: A broad class of statistical models is considered that can be misspecified and ill-posed, from a Bayesian perspective. This provides a flexible and interpretable framework for their analysis, but it is important to understand robustness of the chosen Bayesian model and its effect on the resulting solution, especially in the ill-posed case where in the absence of prior information the solution is not unique.
Compared to earlier work about the Bernstein-von Mises theorem for nonregular well-posed Bayesian models, it is shown that non-identifiable part of the likelihood, together with the constraints on the parameter space, introduce a more complex geometric structure of the posterior distribution around the best reconstruction point in the limit, and a local approximation of the posterior distribution in this neighbourhood is provided.
The results apply to misspecified models which allows, for instance, to evaluate the effect of model approximation on statistical inference. Emission tomography is taken as a canonical example for study, but our results hold for a wider class of generalised linear inverse problems with constraints.

Invited Session IS-64 • Friday 15, 14:30-16:30 • Room Vietri

NEW DEVELOPMENT IN SEMIPARAMETRIC METHODS IN MICROECONOMETRICS AND FINANCE

Organizer(s): Valentina Corradi (University of Surrey)
Chair: Valentina Corradi (University of Surrey)

Title: ***Testing Quantile Sample Selection***
Speaker: **Daniel Gutknecht**, University of Mannheim, Germany

Co-author(s): Valentina Corradi
Schedule: *Friday 15 June, 14:30–15:00*
Abstract: The objective of this paper is to provide a unified framework for testing selection in conditional mean as well as in conditional quantiles. Our testing strategy consists in two steps. In the first step we provide a test for omitted predictors, where the omitted predictor is the propensity score, which can be (non-)parametrically estimated in a previous stage. We show that propensity score estimation error is asymptotically negligible. The test has power against generic $\sqrt{n}$-local alternative, where n denotes the sample size. If there is no endogenous selection, we do not reject the null, if instead there is endogenous selection, we reject it. Hence, if we fail to reject the null, we can confidently conclude that there is no selection. On the other hand, a rejection can be due to either endogenous selection or to omission of a predictor which is not independent of the propensity score. In the second step, we therefore propose a test for conditional (quantile) independence, which ensures that the null is rejected if and only if we have endogenous selection.

Title: ***Intraday vs overnight betas***
Speaker: **Walter Distaso**, Imperial College London, United Kingdom
Co-author(s): Alessandra Insana
Schedule: *Friday 15 June, 15:00–15:30*
Abstract: We uncover substantial variation between intraday and overnight betas. Dividing the total daily return in intraday and overnight, we evaluate intraday and overnight betas. Starting by the classical Capital Asset Pricing Model (CAPM), we estimate our three unconditional betas. We then consider conditional betas using nonparametric methods.
Finally, we evaluate the predictive power of intraday and overnight betas adapting the betting against the betas of Frazzini and Pedersen.

Title: ***A Semiparametric Network Formation Model with Multiple Linear Fixed Effects***
Speaker: **Luis Candelaria**, University of Warwick
Schedule: *Friday 15 June, 15:30–16:00*
Abstract: This paper analyzes a semiparametric model of network formation in the presence of multiple, unobserved, and agent-specific fixed effects. Given agents' observed attributes, the conditional distributions of these effects, as well as the disturbance terms associated with each linking decision are not parametrically specified. I give sufficient conditions for point identification of the coefficients on the observed covariates. This result relies on the existence of at least one

continuous covariate with unbounded support. I provide partial identification results when all covariates have a bounded support. Specifically, I derive bounds for each component of the vector of parameters when all the covariates have a discrete support. I propose a semiparametric estimator for the vector of coefficients that is consistent and asymptotically normal as the number of individuals in the network increases. Monte Carlo experiments demonstrate that the estimator performs well in finite samples. Finally, in an empirical study, I analyze the determinants of a friendship network using the Add Health dataset.

Title: ***Nonparametric Instrumental Variable Estimation of Additive Models***
Speaker: **Sorawoot Srisuma**, University of Surrey, United Kingdom
Co-author(s): Samuele Centorrino
Schedule: *Friday 15 June, 16:00–16:30*
Abstract: We propose a two-step estimator for nonparametric additive regression functions with multiple endogenous and exogenous conditioning variables. In the first step we construct a sieve nonparametric instrumental variable estimator that achieves the optimal rate of convergence in a minimax sense. We smooth this over in the second step using kernel methods. The subsequent estimator has an asymptotic normal distribution and has an oracle property. In particular, the asymptotic distribution of each additive component is the same as it would be if all the other components were known.

Invited Session IS-65 • Friday 15, 14:30-16:30 • Room Ravello

COMPLEX TIME SERIES MODELLING

Organizer(s): Qiwei Yao (London School of Economics)
Chair: Tobias Kley (Humboldt-Universität zu Berlin)

Title: ***Sequential detection of structural changes in irregularly observed data***
Speaker: **Tobias Kley**, Humboldt-Universität zu Berlin, Germany
Co-author(s): Piotr Fryzlewicz
Schedule: *Friday 15 June, 14:30–15:00*
Abstract: Online surveillance of time series is traditionally done with the aim to identify changes in the marginal distribution under the assumption that the data between change-points is stationary and that new data is observed at constant frequency. In many situations of interest

to data analysts, the classical approach can be too restrictive to be used unmodified. We propose a unified system for the monitoring of structural changes in streams of data where we use generalised likelihood ratio-type statistics in the sequential testing problem, obtaining the flexibility to account for the various types of changes that are practically relevant (such as, for example, changes in the trend of the mean). Our method is applicable to sequences where new observations are allowed to arrive irregularly. Early identification of changes in the trend of financial data can assist to make trading more profitably. In an empirical illustration we apply the procedure to intra-day prices of components of the NASDAQ-100 stock market index.

Title: ***Testing different structures of Spatial Dynamic Panel Data models by a bootstrap multiple testing procedure***
Speaker: **Maria Lucia Parrella**, University of Salerno, Italy
Co-author(s): Francesco Giordano, Massimo Pacella
Schedule: *Friday 15 June, 15:00–15:30*
Abstract: In the econometric field, spatio-temporal data is often modeled by means of spatial dynamic panel data models (SDPD). In the last decade, several versions of the SDPD model have been proposed, each one based on different assumptions on the spatial parameters and different properties of the estimators. In particular, the classic version of the model is the one that assumes the spatial parameters to be homogeneous over location. Another version of the model, proposed recently and called Generalized SDPD, assumes that spatial parameters are adaptive over location. In this work we propose a strategy for testing the particular structure of the spatial dynamic panel data model, by means of a multiple testing procedure that allow to choose between the generalized version of the model and some specific versions derived from the general one by imposing particular constraints on the parameters. The theoretical derivations of the testing procedure are made in the high dimensional setup, where the number of locations may grow to infinity with the time series length. This makes our proposal also a nonstandard application of the multiple testing approach, since the dimension of the multiple testing scheme grows to infinite with the sample size.

Title: ***Detecting multiple local extrema via wild binary segmentation***
Speaker: **Yining Chen**, London School of Economics and Political Science, United Kingdom
Co-author(s): Piotr Fryzlewicz
Schedule: *Friday 15 June, 15:30–16:00*
Abstract: We consider the problem of detecting the number and locations of

multiple local maxima and minima given a sequence of univariate observations, in both the nonparametric regression and time series settings. We propose a new approach that combines the ideas of wild binary segmentation (Fryzlewicz, 2014) and mode estimation using isotone regression. We show that our procedure consistently estimates the number of local extrema, and is minimax optimal (up to a logarithmic factor) in estimating the locations of these points. Moreover, we show that the computational complexity of our method is linear up to a logarithmic factor. Finally, we discuss how our approach could be extended to detect other interesting features.

Title: ***The Prediction of CPI in China — Based on Semi-parametric ECM-MIDAS Model***

Speaker: **Dong Yang**, Southwestern University of Finance and Economics, China, People's Republic of

Co-author(s): Wanbo Lu

Schedule: *Friday 15 June, 16:00–16:30*

Abstract: According to the MIDAS model, a semi-parametric ECM-MIDAS model is constructed for the data which exist cointegration relation. By using the generalized likelihood ratio test, the problem of the consistency test of the functional form of the parametric regression model is solved. By using this model, China's stock market weekly data and monthly crude oil market international data are used to make short-term forecasts of China's CPI. Based on the AIC criterion, a comprehensive comparison was made among the four kinds of mixed frequency sampling models which includes the semi-parametric ECM-MIDAS models and the two kinds of same frequency models. The results show that the error correction term has obvious nonlinear characteristics. No matter we use recursive, rolling or fixed sample, the semi-parametric ECM-MIDAS model proposed in this paper has the best prediction accuracy when we make continuous prediction. In addition, the prediction results are not affected by the choice of dynamic mixed frequency cointegrating relationship.

Invited Session IS-66 • Friday 15, 14:30-16:30 • Room Furore

APPLIED MATHEMATICS MEETS STATISTICS AND PROBABILITY

Organizer(s): Anirban Dasgupta (Purdue University)

Chair: Anirban Dasgupta (Purdue University)

Title: ***Optimal link prediction with matrix logistic regression***

Speaker: **Nicolai Baldin**, University of Cambridge, United Kingdom
Co-author(s): Quentin Berthet
Schedule: *Friday 15 June, 14:30–15:00*
Abstract: We consider the problem of link prediction, based on partial observation of a large network, and on side information associated to its vertices. The generative model is formulated as a matrix logistic regression. The performance of the model is analysed in a high-dimensional regime under a structural assumption. The minimax rate for the Frobenius-norm risk is established and a combinatorial estimator based on the penalised maximum likelihood approach is shown to achieve it. Furthermore, it is shown that this rate cannot be attained by any (randomised) algorithm computable in polynomial time under a computational complexity assumption.

Title: ***Obstacle problems for a class of nonlocal operators arising in mathematical finance***
Speaker: **Donatella Danielli**, Purdue University, United States of America
Co-author(s): Arshak Petrosyan, Camelia Pop
Schedule: *Friday 15 June, 15:00–15:30*
Abstract: In this talk we will discuss existence, uniqueness, and regularity of viscosity solutions to the stationary and evolution obstacle problems defined by a class of nonlocal operators that are not stable-like and may have supercritical drift. The class of nonlocal operators that we consider include non-Gaussian asset price models widely used in mathematical finance, such as Variance Gamma Processes and Regular L'evy Processes of Exponential type. In this context, the viscosity solutions that we analyze coincide with the prices of perpetual and finite expiry American options.

Title: ***Generalized CP for variable selection***
Speaker: **Linda Zhao**, University of Pennsylvania, United States of America
Co-author(s): Lawrence Brown, Junhui Cai
Schedule: *Friday 15 June, 15:30–16:00*
Abstract: Linear models as working models have performed very well in practice. But most often the theoretical properties are obtained under the usual linear model assumptions such as linearity, homoscedasticity and normality. Using the least squared estimators we justify their desirable properties under much broader model assumptions, namely a model lean framework. Generalized CP (GCP) is proposed to estimate the prediction errors. We study its properties. An alternative bootstrap method is also investigated. Model selections are done through both methods.
Joint work with Lawrence Brown, Junhui (Jeff) Cai and the Wharton

group

Title: ***Simultaneous non-parametric regression in RADWT dictionaries***
Speaker: **Daniela De Canditiis**, C.N.R., Italy
Co-author(s): Italia De Feis
Schedule: *Friday 15 June, 16:00–16:30*
Abstract: We propose a new technique for non-parametric regression of multichannel signals. The proposed technique fuses together two results, one from the signal processing field and the other from the modern high-dimensional data analysis. The technique is based on the use of the Rational-Dilation Wavelet Transforms (RADWT), equipped with a tunable Q-factor able to provide sparse representations of functions with different oscillations persistence. In particular, we consider two different frames obtained by two RADWT with different Q-factors able to provide sparse representations of functions with low and high resonance. The proposed methodology is based on the hypothesis that the signals are measured simultaneously on several independent channels and that they share the low resonance component and the spectral characteristics of the high resonance component. The regression analysis is then performed by the means of the grouped lasso penalty. Moreover, we present a result of optimality for the norm of the difference between the unknown signals and their estimators. Numerical experiments show the performance of the proposed method in different synthetic scenarios as well as in a real case example for the analysis and joint detection of sleep spindles and K-complex events for multiple EEG signals.

Invited Session IS-67 • Friday 15, 14:30-16:30 • Room Positano

NON-STANDARD PROBLEMS IN MODERN GOODNESS OF FIT THEORY
Organizer(s): Estate Khmaladze (University of Wellington)
Chair: Estate Khmaladze (University of Wellington)

Title: ***Goodness of fit testing for copulas: a distribution-free approach***
Speaker: **Sami Umut Can**, University of Amsterdam
Co-author(s): John H.J. Einmahl, Roger J.A. Laeven
Schedule: *Friday 15 June, 14:30–15:00*
Abstract: Consider a random sample from a continuous multivariate distribution function F with copula C. In order to test the null hypothesis that C belongs to a certain parametric family, we construct an under H_0 asymptotically distribution-free process that serves as a tests genera-

tor. The process is a transformation of the difference of two estimators of C. We show that this transformed empirical process converges weakly to a standard multivariate Wiener process, paving the way for a multitude of asymptotically distribution-free goodness of fifit tests for copula families. We demonstrate through a Monte Carlo simulation study that our approach has excellent fifinite sample performance and illustrate its applicability with a data analysis.

Title: ***Tail dependence structure: Did it change?***
Speaker: **Roger J. A. Laeven**, University of Amsterdam, Department of Quantitative Economics
Co-author(s): S. Umut Can, John H. J. Einmahl
Schedule: *Friday 15 June, 15:00–15:30*
Abstract: Consider two i.i.d. samples independently generated from d-variate distribution functions F and F'. Suppose the two distribution functions lie in the max-domains of attraction of extreme value distributions G and G', and they have associated tail copulas R and R', respectively. We develop a procedure for constructing asymptotically distribution-free tests for the equality of R and R'. We analyze the finite sample behavior of our approach in Monte Carlo simulations. We apply our approach to several financial data-sets.

Title: ***Distribution free GOF tests for some non-parametric hypothesis testing problems***
Speaker: **Thuong Nguyen**, Victoria University of Wellington, New Zealand
Schedule: *Friday 15 June, 15:30–16:00*
Abstract: We consider two typical non-parametric hypothesis testing problems: testing the independence of 2 random variables and testing whether two samples arise from the same distribution, both are in discrete scenarios. For these two problems, there have been so far only one asymptotically distribution-free GOF test available, the chi-square test. The fact is, when the number of outcomes is large and the sample size is relatively modest, the chi-square test is no longer usable. We will talk about a new approach to constructing a whole class of asymptotically distribution-free GOF tests for these two problems. The possibility of these new tests overcome the obstacle with "sparse" data will also be discussed.

Title: ***Asymptotic hypothesis testing for the color-blind problem***
Speaker: **Laura Eliza Dumitrescu**, Victoria University of Wellington, New Zealand
Co-author(s): Estate Khmaladze
Schedule: *Friday 15 June, 16:00–16:30*

Abstract: In the non-parametric framework, the conventional approach for testing distributions equality (to Q) in the classical two-sample problem, is based on the difference between the two marginal empirical distributions. When the order within each pair is unobservable, it is no longer possible to compute these and an approach based on a symmetrised empirical process is proposed. Based on its form, linear test statistics are derived, which have distributions independent of the (unspecified) Q, and are optimal for particular local alternatives.

Invited Session IS-68 • Friday 15, 14:30-16:30 • Room Maiori

OPTIMAL TRANSPORT FOR NON PARAMETRIC STATISTICS

Organizer(s): Jean-Michel Loubes (Univeristy of Toulouse)

Chair: Jean-Michel Loubes (Univeristy of Toulouse)

Title: ***A Gaussian Process Regression Model for Distribution Inputs***

Speaker: **Nil Venet**, Università degli studi di Bergamo, Italy

Co-author(s): François Bachoc, Fabrice Gamboa, Jean-Michel Loubes

Schedule: *Friday 15 June, 14:30–15:00*

Abstract: Monge-Kantorovich distances, otherwise known as Wasserstein distances, have received a growing attention in statistics and machine learning as a powerful discrepancy measure for probability distributions. We focus on forecasting a Gaussian process indexed by probability distributions. For this, we provide a family of positive definite kernels built using transportation based distances. We provide a probabilistic understanding of these kernels and characterize the corresponding stochastic processes. We prove that the Gaussian processes indexed by distributions corresponding to these kernels can be efficiently forecast, opening new perspectives in Gaussian process modeling.

Title: ***Central Limit Theorem for Wasserstein Barycenters of Gaussian Measures***

Speaker: **Aleksei Kroshnin**, Institute for Information Transmission Problems, Russian Federation; Higher School of Economics, Russian Federation; Moscow Institute of Physics and Technology, Russian Federation

Co-author(s): Alexandra Suvorikova

Schedule: *Friday 15 June, 15:00–15:30*

Abstract: Optimal transportation metrics, e.g. 2-Wasserstein metricW_2, is a

natural way to measure distance between probability distributions on some space X taking into account the underlying geometry of X. Moreover, it gives a possibility to define a nonlinear, "geometrical" averaging of measures via Frechet mean. Namely, an average of measures $\mu_1, \ldots, \mu_n \in P(X)$ called Wasserstein barycenter is a measure which minimizes $\frac{1}{n}\sum_{i=1}^{n} W_2^2(\mu_i, \nu)$ over $\nu \in P(X)$. In stochastic setting measures are independently drawn from some distribution on $P(X)$. It is known that the Law of Large Numbers holds for Wasserstein barycenters in quite a general situation, i.e. empirical barycenter of first n measures converges to the barycenter of distribution in Wassestein distance. However, the Central Limit Theorem is still an open question. The work presents the CLT for barycenters of Gaussian measures: we show that it holds without any additional assumptions and establish some quantitative bounds for limit distribution.

Title: ***Gaussian process forecast with multidimensional distributional input***

Speaker: **Alexandra Suvorikova**, WIAS, Germany

Co-author(s): Francois Bachoc, Jean-Michel Loubes, Vladimir Spokoiny

Schedule: *Friday 15 June, 15:30–16:00*

Abstract: In this work, we focus on forecasting a Gaussian process indexed by elements in 2-Wasserstein space. We introduce a family of positive definite kernels constructed with the use of L_2-distance between optimal transportation maps and provide their probabilistic understanding. The technique allows to forecast efficiently Gaussian processes, which opens new perspective in Gaussian process modeling.

Title: ***Histograms registration/warping using regularized barycenters in the Wasserstein space***

Speaker: **Elsa Cazelles**, Institut de Mathématiques de Bordeaux, Université de Bordeaux, France

Co-author(s): Jérémie Bigot, Nicolas Papadakis

Schedule: *Friday 15 June, 16:00–16:30*

Abstract: We present a framework to align data in the form of (possibly noisy) histograms in d-dimensional euclidean space. Such work takes interest in biology, where researchers aim to normalize automatically large datasets in order to compare characteristics within a same cell population, while dealing with noise in sample acquisition due to technical variations of the environment. To that end, we propose to compute a Fréchet mean (or barycenter) of a sample set with respect to the Wasserstein metric on probability space. In order to

handle aberrant data and to remove acquisition effects, we study two different regularizations of the Fréchet mean. The first one consists in penalizing the barycenter with a convex functional [Penalized barycenters in the Wasserstein space. J. Bigot, E. Cazelles, N. Papadakis]. The second scheme modifies the Wasserstein distance itself [Sinkhorn Distances: Lightspeed Computation of Optimal Transport. M. Cuturi] by adding an entropy regularization. We then propose to automatically set the parameter of each method using Lepskii's data-driven principle. Both approaches are illustrated with simulated gaussian mixtures and flow cytometry data.

Invited Session IS-69 • Friday 15, 14:30-16:30 • Room Procida

VARIOUS TOPICS

Organizer(s): Anton Schick (Binghamton University)
Chair: Anton Schick (Binghamton University)

Title: ***Estimating the response density in semiparametric regression***
Speaker: **Ursula U. Müller**, Texas A&M University, United States of America
Schedule: *Friday 15 June, 14:30–15:00*
Abstract: This talk focuses on regression models with a parametric (linear or nonlinear) regression function. We assume that the errors have mean zero and are independent of the covariates. The independence assumption is important: it enables us to construct an estimator for the response density that uses all the observed data, in contrast to the usual local smoothing techniques, and which therefore permits a faster rate of convergence. For invertible (monotonic) regression functions, and a suitably chosen bandwidth, this estimator is consistent and converges with the optimal parametric root-n rate. We also discuss cases when the regression function is not invertible and the root-n rate cannot be achieved. If the regression function is a step function, we provide a response density estimator that has the same bias as the usual estimators based on the responses, but a smaller asymptotic variance.

Title: ***Plug-in estimators for random fields with nearest neighbor interactions***
Speaker: **Wolfgang Wefelmeyer**, University of Cologne, Germany
Schedule: *Friday 15 June, 15:00–15:30*
Abstract: Consider a random field with nearest neighbor interactions on a d-dimensional square lattice. We show that the empirical estimator for

the expectation of a local function on the field is efficient if and only if this function is a sum of functions each of which depends only on the values of the field on a clique of sites. For local functions involving larger configurations than cliques, we construct estimators with smaller asymptotic variance than the empirical estimator as follows. First, using splitting properties of the field, we factor its distribution into conditional distributions involving at most $d+1$ sites. Then we write the expectation of a local function in terms of these conditional distributions. For countable state space, we estimate conditional probabilities using empirical probabilities. If the state space is a subset of the real line, and the field has a sufficiently smooth density, we estimate conditional densities using appropriate kernel density estimators. This is joint work with Cindy Greenwood and Ian McKeague.

Title: ***Semi-Parametric Modeling of Structured Point Processes Using Multi-Level Log-Gaussian Cox Processes***

Speaker: **Ganggang Xu**, Binghamton University (SUNY), United States of America

Co-author(s): Ming Wang, Hui Huang, Jingfei Zhang, Yongtao Guan

Schedule: *Friday 15 June, 15:30–16:00*

Abstract: We propose a general framework of using multi-level log-Gaussian Cox processes to model repeatedly observed point processes with complex structures. A novel nonparametric approach is developed to consistently estimate the covariance kernels of the latent Gaussian processes at all levels. Consequently, multi-level functional principal component analysis can be conducted to investigate the various sources of variations in the observed point patterns. In particular, to predict the functional principal component scores, we propose a consistent estimation procedure by maximizing the conditional likelihoods of super-positions of point processes. We further extend our procedure to the bivariate point process case in which potential correlations between the processes can be assessed. Asymptotic properties of proposed estimators are investigated, and the effectiveness of our procedures is illustrated through a simulation study and an application to a stock trading dataset.

Title: ***Estimation of the error distribution function in a varying coefficient regression model***

Speaker: **Anton Schick**, Binghamton University, United States of America

Co-author(s): Yilin Zhu, Xiaojie Du

Schedule: *Friday 15 June, 16:00–16:30*

Abstract: This talk discusses estimation of the error distribution function

in a varying coefficient regression model. Three estimators are introduced and their asymptotic properties described by uniform stochastic expansions. The first estimator is a residual-based empirical distribution function utilizing an under-smoothed local quadratic smoother of the coefficient function. The second estimator exploits the fact that the error distribution has mean zero. It improves on the first estimator, but is not yet efficient. An efficient estimator is obtained by adding a stochastic correction term to the second estimator.

Authors index