ISNPS 2022

Programme

WITH ABSTRACTS

20 - 24 JUNE 2022

MONDAY 20 JUNE 2022

8:45 - 9:00 Welcome

9:00 - 10:00 Peter Hall's Lecture: Steve Marron

Chair: Regina Liu
Room: Akamas A

Peter Hall and High Dimension Low Sample Size Asymptotics

Abstract: Personal memories of Peter Hall, together with a research talk on an area that we founded together: High Dimension Low Sample Size Asymptotics. This branch of research is very relevant for modern high dimensional data, and is a clear example how it is both essential, and also beautiful, for statistical intuition to be guided by mathematics.

10:00 - 11:00 Contributed Paper Session 1

Bayesian inference

Chair: Matteo Giordano
Room: Akamas C

10:00 **A Bayesian Nonparametric Approach To Super-Resolution Single-Molecule Localization**
Mariano Gabitto, Michael Jordan, Herve Marie-Nelly, Xavier Darzacq, Ari Pakman

Abstract: We consider the problem of single-molecule identification in super-resolution microscopy. Super-resolution microscopy overcomes the diffraction limit by localizing individual fluorescing molecules in a field of view. This is particularly dicult since each individual molecule appears and disappears randomly across time and because the total number of molecules in the eld of view is unknown. Additionally, data sets acquired with super-resolution microscopes can contain a large number of spurious fluorescent fluctuations caused by background noise. To address these problems, we present a Bayesian nonparametric framework capable of identifying individual emitting molecules in super-resolved time series. We tackle the localization problem in the case in which each individual molecule is already localized in space. First, we collapse observations in time and develop a fast algorithm that builds upon the Dirichlet process. Next, we augment the model to account for the temporal aspect of Fluorophore photo-physics. Finally, we assess the performance of our methods with ground-truth data sets having known biological structure.

10:20 **Nonparametric Bayesian inference for reversible multi-dimensional diffusions**
Matteo Giordano

Abstract: Reversible multi-dimensional diffusion processes are ubiquitous models for the motion of particles diffusing in a potential energy field. The talk will consider the problem of estimating the potential function from an observed continuous trajectory. A nonparametric Bayesian approach is employed based on Gaussian and p-exponential priors. The posterior distributions arising from both classes of priors are shown to attain, as the observation time increases, optimal posterior contraction rates towards the ground truth in any dimension. These results are based on a general posterior contraction rate theorem that exploit reversibility to construct suitable tests for the invariant measure.

Survival analysis I

Chair: Ali Shariati
Room: Aphrodite A

10:00 **Quantile Regression for Interval-Censored Data using Laguerre Polynomials**
Benjamin Deketelaere, Ingrid Van Keilegom, Anouar El Ghouch

Abstract: Quantile regression(QR) is an alternative to the common mean regression. In QR, conditional quantiles of a response variable are estimated across values of the predictor variables. In the standard setting of complete observed data, the regression coefficients are estimated by minimizing a check loss function. In this work we consider QR in the case of interval-censored data. Minimizing the standard check loss function is no longer possible with interval-censored observations. As a first step, it is possible to transform the minimization of the check function into a "maximum likelihood problem". Indeed, if we assume that the response variable conditional to the explanatory variables follows an Asymetric Laplace distribution(ALD), it can be shown that minimizing the check function is equivalent to fit an ALD to the data. For censored data, fitting an ALD leads to inconsistent estimation of the regression coefficients. We propose to "enrich" the Laplace Distribution by using Laguerre polynomial expansions. The idea behind this is to use the extra-flexibility provided by the Laguerre polynomials in order to better approximate the underlying distribution. The outline of the presentation is as follows. To begin, the methodology will be introduced and motivations for the use of Laguerre polynomials will be provided. Then, the consistence and a convergence rate for our estimator will be discussed. Also, the results of some finite samples simulations are shown. These demonstrate that using the Enriched Laplace Distribution leads to a better bias reduction than other existing estimating methods. Finally, an application on a real dataset will be presented.

10:20 **Challenges with model and tuning parameter selection when using cross-validation with censored data**
Anders Munch

Abstract: Most machine learning algorithms depend on one or more tuning parameters to control the trade-off between bias and variance. To select the optimal value for a tuning parameter, a popular approach is to use cross- validation to determine which value minimizes a given loss function. We consider the problem of selecting a regression model from a collection of candidate models using cross-validation on an external data set where the observations might be right-censored in continuous time. In this situation it is common to use the component of the negative log-likelihood corresponding to the outcome as the loss function. This ignores the contribution to the likelihood made by the censoring distribution, and we argue that this approach is problematic in at least two ways: Firstly, we show that the least false parameter according to this loss function is in general not well-defined as it depends on the censoring distribution. Secondly, for many commonly used survival models the likelihood will a.s. be zero for any hold-out sample. This means that the negative log- likelihood loss cannot be used to compare general survival models and hence does not provide a general approach for model selection in the survival setting. We discuss how these problems can be alleviated by modeling the censoring distribution, which, on the other hand, comes at the cost of introducing a new nuisance parameter to be estimated.

10:40 **Asymptotic Behaviour of the MRL Function Estimator with Length-biased and Right Censored Data**
Ali Shariati, Hassan Doosti, Vahid Fakoor

Abstract: Prevalent cases recruited through a cross-sectional sampling procedure may be followed over time to assess the progress of a specific disease. Right censored survival data collected in such studies is not statistically representative of the target population of interest due to nonrandom sampling of the subjects. It is well known that when the incidence is constant, the observed sample is length-biased. It is worth noting that the Kaplan–Meier estimator is not the nonparametric MLE of the length-biased distribution function owing to informative censoring induced by the sampling mechanism. When communicating with non-statisticians, average remaining lifetime is a more meaningful and comprehensible measure than the survival probability or the hazard rate. Therefore, this talk is centered on the mean residual lifetime (MRL) function. The nonparametric MLE of the unbiased MRL function is introduced using length-biased and right censored data. The asymptotic properties of the nonparametric MLE is then discussed. These results are employed to derive confidence bands and intervals for the MRL function. A simulation study is conducted to examine the finite sample performance of the proposed nonparametric estimator. The nonparametric MLE and the asymptotic properties are then applied to obtain the MRL and the respective confidence bands/intervals for elderly residents of a retirement centre.

Robust statistics I

Chair: Alexandre Lecestre
Room: Aphrodite B

10:00 **A Data-Driven Strategy for Specifying Cutoffs in Trimming and Winsorizing**
Derek Young, Kedai Cheng

Abstract: Trimming and Winsorizing are classic approaches for developing robust statistics. One issue with these approaches is how to specify the cutoff level. Many strategies have been developed in the context of survey data analysis and assume the availability of previous information; e.g., population estimates from a survey. We propose a data-driven approach when such previous information is not available or when the researcher is not willing to make distributional assumptions. Our approach utilizes the intrinsic meaning of tolerance intervals, which bound a specified proportion of the population at a given confidence level. In the multivariate context, this involves ordering the data using statistical data depth, finding the depth value that corresponds to a nonparametric one-sided lower tolerance limit, then constructing the convex hull of the set of all points with a depth value greater than that limit. The amount of trimming or Winsorizing can now be guided by the tolerance interval's levels, and the resulting convex region can be used as the basis for performing robust inference. In this talk, we present the full methodological development of what is proposed above and highlight the trimmed and Winsorized means based on this cutoff strategy. We also study the influence function of these estimators and present their limiting distributions, which can be derived via their asymptotic representations. Extensive numerical results show the efficacy of and improved interpretation when taking the proposed data- driven strategy compared to classic pre-specified cutoff levels. Our approach is demonstrated on farmland data from the quinquennial Census of Agriculture, which is conducted by the U. S. Department of Agriculture's National Agricultural Statistics Service. Using our cutoff strategy, robust bivariate means are constructed starting with data from the 1982 Census of Agriculture, which we interpret across the eight censuses analyzed.

10:20 **Robust estimation under a shape contraint**
Hélène Halconruy, Yannick Baraud, Guillaume Maillard

Abstract: The problem of estimating a density under a shape contraint has been widely investigated since a seminal paper by Grenander (1956). To estimate a nonincreasing density on the half line, Grenander designed an eponymous estimator that coincides with the maximum likelihood estimator (MLE) over the set of nonincreasing densities on half line. There, the MLE exhibits the surprising property of adapting its convergence rate to the -a priori unknown- specific features of the target density. MLE's adaptation property in this case has made it the favourite and almost exclusive estimator to tackle other shape-constrained estimation problems such as convexity, k-monotonicity, log-concavity (in higher dimension)… However, the MLE has some drawbacks; first, it requires knowledge of certain information about the target density support possibly unknown in practice. Moreover, it may perform poorly when data are only close to being i.i.d. In a joint work with Y. Baraud and G. Maillard, we design in the one dimensional case, an estimator that keeps MLE's

minimax and adaptation properties for shape-constrained density estimation and that is robust i.e., that remains stable with respect to a slight deviation from the ideal situation of truly i.i.d. data whose density satisfies the required shape. In particular our estimator still performs when: the condition of equidistribution is slightly violated, there exists a small portion of outliers in the sample, the data true density (equidistributed case) does not satisfy the shape constraint but is close enough (for some loss) to a density that does. In this talk, I will present our procedure based on the l-estimation (Baraud, 2021), give its risk bounds for the total variation distance and finally illustrate it in shape-constrained estimation problems such as monotonicity on a half line and convexity on an interval.

10:40 **Robust estimation in finite state space hidden Markov models**
Alexandre Lecestre

Abstract: We consider stationary hidden Markov models with finite state space and estimate the stationary law of consecutive observations with a focus on robustness properties of our estimator. This means we do not assume the process to be exactly a hidden Markov chain nor to be exactly stationary and we consider the possible presence of outliers. We proved a non-asymptotic sample bound on the resulting squared Hellinger error when the emission densities of our model belong to exponential families. We can derive convergence rates assuming a well specified setting a posteriori, those rates are optimal up to logarithmic factors. It is possible to derive a bound for the estimation of parameters with additional assumptions.

Clustered data

Chair: Erin Sprünken

Room: Christian Barnard

10:00 **Leave-cluster-out technique and variance estimation when regressors are many**
Stanislav Anatolyev

Abstract: We introduce the leave-cluster-out (LCO) machinery for clustered samples, a generalization of leave-one-out methods that prove useful for independent data. We use LCO to construct an estimator of the asymptotic variance of the OLS estimator in a linear regression characterized by possibly numerous regressors and arbitrary within-cluster heteroskedasticity. We show consistency of the LCO variance estimator when regressors may be many, regression errors may be heteroskedastic, clusters may be unbalanced and heterogeneous, and cluster sizes may be moderately large. Simulations reveal amazing robustness of the LCO estimator to regressor numerosity and heteroskedasticity.

10:40 **A generalized framework for rank-based analysis of cluster data in the several sample case**
Erin Sprünken

Abstract: In many trials and experiments, subjects are not only observed once, but multiple times, resulting in a cluster of possibly correlated observations. For example, mice sharing the same cage or students of the same class are typical examples of clustered data. Typically, under the assumption of normally distributed data, mixed models are used for analysis. However, this model assumption is rather strict and hard to justify in most real data analyses. Furthermore, skewed data (e.g. waiting times), discrete data (e.g. count data) or ordered categorical data measured on an ordinal scale are typical endpoints in a variety of trials. This motivates the use of nonparametric methods which do not rely on any specific data distribution.

For the two-sample case, several nonparametric procedures exist. For binary clustered data, a chi-square-test for contingency tables can be used. Furthermore, generalizations of the Wilcoxon-Mann-Whitney-test exist for testing the null hypothesis of equal distributions of clustered data. An extension is provided by a procedure under a less strict null hypothesis formulated in terms of the Wilcoxon-Mann-Whitney effect. In the present talk, we aim to generalize the procedures for the analysis of several samples. Thus, we propose a general nonparametric framework for comparing multiple groups of clustered data under mild assumptions. We present different inference methods, namely ANOVA-type test statistics and a multiple contrast test procedure and investigate their asymptotic behavior. Extensive simulation studies indicate that the methods control the type-1 error level well, even with small sample sizes. A real data example illustrates the application of the proposed methods.

Model selection

Chair: Ulrike Schneider

Room: Leda

10:00 **Model selection for robust learning of mutational signatures under overdispersion**
Marta Pelizzola, Ragnhild Laursen, Asger Hobolth

Abstract: Mutational signatures are derived from a collection of mutational profiles using non-negative matrix factorization (NMF). To extract the mutational signatures we have to determine an error model for the observed mutational counts and a number of mutational signatures. In most applications, the mutational counts are assumed to be Poisson distributed, but they are often overdispersed. In the case of overdispersion, the Negative Binomial distribution is more appropriate. We introduce a procedure where the error model is determined from residual fits and the number of signatures is chosen using cross-validation. We demonstrate using a simulation study that Poisson NMF often leads to an overestimation of the number of signatures as a way to account for the overdispersion in the data. In the simulation study, we also compare our cross-validation procedure to two classical model selection procedures: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). We find that AIC tends to overestimate and BIC tends to underestimate the number of signatures, whereas our cross-validation procedure finds the right balance between the fit to the data and the complexity

of the model. Furthermore, we show that our novel cross-validation procedure for model selection is less influenced by a wrong distributional assumption compared to other commonly used methods for extracting mutational signatures.

10:20 **The Probability of Improved Prediction: a new concept in statistical inference**
Stijn Jaspers, Olivier Thas

Abstarct: In many empirical sciences there is a consensus that the Scientific Method can rely on statistical hypothesis testing and that the p-value can be compared to a threshold of e.g. 5% to come to a binary decision: either reject the null hypothesis (a "positive" result) or accept the null hypothesis (inconclusive). The former leads easier to publication, whereas many journals are hesitant publishing negative results. This tradition leads to publication bias. Moreover, sometimes researchers do significance-fishing, which results in an increase of the false positive rate and in irreproducible results. Many scientists and statisticians are criticising this use of p-values nowadays. We propose an alternative summary of experimental data that may overcome some of the issues related to p-values. We believe that our statistic is relevant to the Scientific Method, because it has an interpretation in terms of predictability of a model. In this sense, it may help to connect statistical science with predictive modelling or machine learning. More specifically, we propose the Probability of Improved Prediction (PIP), which measures how more often a model gives better predictions than another model. In this talk, we will focus on two models that differ in a single predictor to compare the performance of the PIP to the p-value in a standard regression model. Firstly, a theoretical comparison is performed in the (unrealistic) scenario where the true model is known. Secondly, it is shown how the new method can be applied in more realistic scenarios and how it performs in these scenarios as compared to the standard approach.

10:40 **On the geometry of uniqueness and model selection of LASSO, SLOPE and related estimators**
Ulrike Schneider, Patrick Tardivel

Abstract: This talk follows the recent trend of exploiting geometric properties in the context of statistical procedures in high-dimensional models. We consider estimation methods such as the Lasso and SLOPE, which are defined as solutions to a penalized optimization problem and provide a geometric condition for uniqueness of the estimator. In contrast to other conditions in the literature, our approach yields a criterion that is both necessary and sufficient. Moreover, these geometric considerations also give insights into which models are accessible for the corresponding estimation method, which we illustrate for the SLOPE estimator using the sign permutahedron.

11: 00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk: Richard Nickl
Chair: Ricardo Cao
Room: Akamas A

Bayesian non-linear inverse problems: progress and challenges

Abstract: Common examples for non-linear inverse problems range from parameter identification in PDEs to tomography and data assimilation problems. They naturally involve high- or infinite dimensional parameter spaces and appropriate statistical noise models lead to a class of non-convex inference problems that present substantial challenges in contemporary data science. In influential work, Andrew Stuart (2010) has proposed a unified Bayesian approach to solve such problems. It is computationally feasible via Gaussian process priors and high-dimensional MCMC algorithms and provides important uncertainty quantification methodology ('error bars' or confidence regions) based on posterior distributions. Despite evident empirical success, the theoretical understanding of the performance of such methods has been limited until recently. Specifically in non-linear settings Bayesian methods are distinct from optimisation based algorithms and their analysis requires a very different set of mathematical ideas. We will summarise recent developments that allow to give rigorous statistical and computational guarantees for the use of these algorithms.

12:30 - 13:30 Lunch Break

13:30 - 15:30 Invited Paper Session 1

Structural inference in high dimensional models
Organiser: Eduard Belitser
Chair: Eduard Belitser
Room: Akamas A

13:30 **Dimension Estimation using Random Connection Models**
Paulo Serra, Michel Mandjes

Abstract: In statistics we often want to discover (sometimes impose) structure on observed data, and dimension plays a crucial role in this task. For instance, high-dimensional data sometimes live in a lower dimensional space; we refer to this as the intrinsic dimension of the dataset. Dimensionality reduction techniques (e.g., PCA, manifold learning) usually rely on knowledge about intrinsic dimension.

Knowledge about dimension is also important to try to avoid the curse of dimensionality. From a computational perspective, the dimension of a dataset has impact in terms of the amount of space needed to store data (compressibility), and the speed of algorithms is also commonly affected by the dimension of input data.The setting considered in this talk is the following: We have access to a certain graph where each vertex represents an observation, and there is an edge between two vertices if the corre- sponding observations are close in some metric. The goal is to estimate the intrinsic dimension of the high-dimensional dataset from this graph only. I give some conditions under which the dimension can be estimated consistently, and some bounds on the probability of correctly recuperating an integer intrinsic dimension. I will also show some numerical results and compare our estimators with some competing approaches from the literature. This is joint work with Michel Mandjes of the University of Amsterdam, the Netherlands.

14:00 **Characterizing the Type 1-Type 2 Error Trade-off for SLOPE**
Cynthia Rush

Abstract: Sorted L1 regularization has been incorporated into many methods for solving high-dimensional statistical estimation problems, including the SLOPE estimator in linear regression. In this talk, we study how this relatively new regularization technique improves variable selection by characterizing the optimal SLOPE trade-off between the false discovery proportion (FDP) and true positive proportion (TPP) or, equivalently, between measures of type I and type II error. Additionally, we show that on any problem instance, SLOPE with a certain regularization sequence outperforms the Lasso, in the sense of having a smaller FDP, larger TPP and smaller L2 estimation risk simultaneously. Our proofs are based on a novel technique that reduces a variational calculus problem to a class of infinite-dimensional convex optimization problems and a very recent result from approximate message passing (AMP) theory. With SLOPE being a particular example, we discuss these results in the context of a general program for systematically deriving exact expressions for the asymptotic risk of estimators that are solutions to a broad class of convex optimization problems via AMP. Collaborators on this work include Zhiqi Bu, Jason Klusowski, and Weijie Su (https://arxiv.org/abs/1907.07502 and https://arxiv.org/abs/2105.13302) and Oliver Feng, Ramji Venkataramanan, and Richard Samworth.

(https://arxiv.org/abs/2105.02180).

14:30 **Optimal Bayesian classification for high dimensional data**
Subhashis Ghoshal

Abstract: Classification of items in one of the two or more given classes based on auxiliary measurements is a fundamental problem of statistical decision making in face of uncertainty. Linear and quadratic discriminant analysis provide optimal model-based classification rules, which require estimation of the precision matrix in a multivariate normal population. Modern data often involve high dimensional measurements, making an accurate estimation of the precision matrix difficult, and hence compromising the accuracy of classification rules. However accurate estimation of a precision matrix in high dimension is possibly under a sparsity assumption that many off-diagonal entries of the precision matrix are zero, which corresponds to conditional independence between the resulting variables given others. We consider a Bayesian approach to classification by inducing sparsity through a shrinkage prior on the Cholesky decomposition of the precision matrix. We show that the posterior for the precision matrix contracts at the optimal rate and the resulting misclassification error of the Bayes classifier converges to that of the oracle Bayes classifier. In simulation studies, we demonstrate the good performance of the proposed Bayesian method. We apply the method to a tumor classification problem. The talk is based on joint work with Xingqi Maggie Du.

15:00 **General framework for projection structures**
Eduard Belitser

Abstract: We develop a general framework for projection structures and study several inference problems within this framework by using the empirical Bayes approach. The proposed general framework unifies a very broad class of high-dimensional models and structures, interesting and important on their own right. We apply the developed theory and demonstrate how the general results deliver a whole avenue of local and global minimax results for particular models and structures as consequences, including white noise model and density estimation with smoothness structure, linear regression and dictionary learning with sparsity structures, bi-clustering and stochastic block models with clustering structure, and others. Various adaptive minimax results over various scales follow also from our local results.

Topics in curve estimation

Organiser: Anton Schick

Chair: Uschi Müller

Room: Akamas C

13:30 **Multivariate, Heteroscedastic Empirical Bayes via Nonparametric Maximum Likelihood**
Bodhisattva Sen, Jake Soloff, Adityanand Guntuboyina

Abstract: Multivariate, heteroscedastic errors complicate statistical inference in many large-scale denoising problems. Empirical Bayes is attractive in such settings, but standard parametric approaches rest on assumptions about the form of the prior distribution which can be hard to justify and which introduce unnecessary tuning parameters. In this talk we extend the nonparametric maximum likelihood estimator (NPMLE) for Gaussian location mixture densities to allow for multivariate, heteroscedastic errors. NPMLEs estimate an arbitrary prior by solving an infinite-dimensional, convex optimization problem; we show that this convex optimization problem can be tractably approximated by a finite-dimensional version. We introduce a dual mixture density whose modes contain the atoms of every NPMLE, and we leverage the dual

both to show non-uniqueness in multivariate settings as well as to construct explicit bounds on the support of the NPMLE. The empirical Bayes posterior means based on an NPMLE have low regret, meaning they closely target the oracle posterior means one would compute with the true prior in hand. We prove an oracle inequality implying that the empirical Bayes estimator performs at nearly the optimal level (up to logarithmic factors) for denoising without prior knowledge. We provide finite-sample bounds on the average Hellinger accuracy of an NPMLE for estimating the marginal densities of the observations. We also demonstrate the adaptive and nearly-optimal properties of NPMLEs for deconvolution. We apply the method to two astronomy datasets, constructing a fully data-driven color-magnitude diagram of 1.4 million stars in the Milky Way and investigating the distribution of chemical abundance ratios for 27 thousand stars in the red clump.

14:00 **Learning high-dimensional functions by conditional least squares**
Stefano Vivogna

Abstract: Estimation on high-dimensional data have become the typical routine of modern machine learning. In this scenario, the statistical accuracy of classical methods can deteriorate dramatically, prompting a race to ever more complex models. On the other hand, many real-world data are inherently structured, hiding low-dimensional relations on which oracle classical models could work just fine. In this talk I will show how simple conditional modifications of super classical methods can learn and adapt to such low-dimensional structures. In particular, statistical optimality in high-dimensions can be restored by conditioning the simplest method --namely, ordinary least squares.

14:30 **Nonparametric inference for general categorical time series with time-varying parameters**
Lionel Truquet

Abstract: The first motivation of this paper is to construct a general framework for modeling non-stationary categorical time series with time-varying analogues of logistic, multiple choice or ordinal time series models and for which some kind of non-stationary exogenous regressors can be included in the dynamic. To this end, we develop two locally stationary notions for autoregressive categorical processes. The first one, which is adapted to strictly exogenous covariates, is based on the theory of Markov chains in random environments and the second one, which is adapted to sequentially exogenous covariates, on iterated random maps systems. In both cases, our results allow to derive asymptotic properties of localized partial sums and consistency properties of local likelihood estimators for time-varying parameters.
For strictly exogenous regressors, we also develop a notion of derivative processes which is useful to control the bias of localized partial sums.

15:00 **Location estimation, empirical Bayes, score matching, and their connection**
Min Xu

Abstract: n this talk, we consider two seemingly unrelated question. The first question is motivated by the following observation. when X_1, ..., X_n ~ Unif(theta_0 - 1, theta_0 + 1), the sample midrange (X_(n) + X_(1))/2 estimates theta_0 with a much smaller error of order 1/n compared with 1/root(n) error rate of the usual sample mean estimator. However, sample midrange is a poor choice when the data has say the N(theta_0, 1) distribution. The natural question then is, can we construct an estimator of theta_0 whose rate is adaptive to the underlying distribution? We propose an estimator based on minimizing L_beta norm where beta is selected in a data driven way. The second question is the vector-valued normal means problem where we observe random vectors Y_i = mu_i + epsilon_i and aim to estimate mu_i. When mu_i has a prior, Tweedie's formula expresses the conditional expectation E[mu_i | Y_i=y] in terms of the multivariate distribution of Y_i. We propose an estimator based on an independent component analysis assumption. Finally, we establish connections between the two problems, showing that they are both specific instances of a general procedure called score matching.

New directions in Bayesian nonparametric modeling

Organiser: Antonio Lijoi

Chair: Antonio Lijoi

Room: Aphrodite A

13:30 **Bayesian nonparametric methods for conditional independence testing**
Sarah Filippi

Abstract: During this talk, I will present Bayesian nonparametric methods for hypothesis testing. In particular I will focus on methods for quantifying the relative evidence in a dataset in favour of the dependence or independence of two variables conditionally on other variables. The approaches use Polya tree priors on spaces of probability densities, accounting for uncertainty in the form of the underlying distributions in a nonparametric way. The Bayesian perspective provides an inherently symmetric probability measure of conditional dependence or independence, a feature particularly advantageous in causal discovery.

14:00 **Exact inference for a class of non-linear hidden Markov models on general state spaces**
Guillaume Kon Kam King, Matteo Ruggiero, Omiros Papaspiliopoulos

Abstract: Filtering hidden Markov models, or sequential Bayesian inference on the hidden state of a signal, is analytically tractable only for a handful of models. Examples are finite-dimensional state space models and linear Gaussian systems (Baum-Welch and Kalman filters). Recently, Papaspiliopoulos et al. ([1], [2]) proposed a principled approach for extending the realm of analytically tractable models, exploiting a duality relation between the hidden process and an auxiliary process. Depending on the dual process, the solution of the filtering problem may consist in a finite or infinite mixture of distributions. In the first case, it is possible to perform exact inference, and we present a study of the computational effort required to implement this strategy for two models: the Cox-Ingersoll-Ross process and the K-dimensional Wright-Fisher process. In particular we propose targeted approximations whose computational complexity is linear in the number of observations [3]. In the second case, we examine Monte-Carlo approximations which allows exploiting duality for a larger class of models.

14:30 **Compound vectors of subordinators and their associated positive Lévy copulas**
Fabrizio Leisen, Alan Riva Palacio

Abstract: Lévy copulas are an important tool which can be used to build dependent Lévy processes. In a classical setting, they have been used to model financial applications. In a Bayesian framework they have been employed to introduce dependent nonparametric priors which allow to model heterogeneous data. This talk focuses on introducing a new class of Lévy copulas based on a class of subordinators recently appeared in the literature, called compound random measures. The well-known Clayton Lévy copula is a special case of this new class. Furthermore, we provide some novel results about the underlying vector of subordinators such as a series representation and relevant moments. This is a work in collaboration with Alan Riva-Palacio.

15:00 **Nonparametric priors for partially exchangeable data: dependence structure and borrowing of information**
Beatrice Franzolini, Antonio Lijoi, *Igor Pruenster*, Giovanni Rebaudo

Abstract: Partial exchangeability is the ideal probabilistic framework for analyzing data from different, though related, sources. The implications on the induced dependence structure and borrowing of information across groups are explored. These findings inspire a new general class of non-parametric priors, termed multivariate species sampling models, which is characterized by its partially exchangeable partition probability function. This class encompasses several popular dependent nonparametric priors and has the merit of highlighting their core distributional properties.

Recent advances in functional and complex data analysis

Organiser: Paromita Dubey & Hanlin Shang

Chair: Philip Reiss

Room: Aphrodite B

13:30 **Functional depth: Recent progress and perspectives**
Stanislav Nagy

Abstract: The depth is a tool of nonparametric statistics. Its objective is to generalise quantiles, rankings, and orderings to multivariate and non-Euclidean data. While a rich body of literature on various depths and depth-like procedures exists, many open problems still stimulate research in the area. We consider the depth of random functions. We revisit the very definition of the standard depths for functional data and introduce procedures allowing adaptive selection of a depth in functional data analysis. Secondly, we draw connections of the functional depth research with topics firmly established in the statistical machine learning literature.

14:00 **Random cohort effects and age groups dependency structure for mortality modelling and forecasting: Mixed-effects time-series model approach**
Bo Wang, Ka Kin Lam

Abstract: Continuous growth in life expectancy has posed significant challenges to many government sectors and life insurance industry due to the extra burdens on the health care services and the rapid increase in pension and insurance expenditure. The problem caused by the ageing population is known as the 'longevity risk'. There have been significant efforts devoted to solving the longevity risk because the continuous growth in population ageing has become a severe issue for many developed countries over the past few decades. We propose a novel mixed-effects time-series approach for mortality modelling and forecasting with considerations of age groups dependence and random cohort effects. The proposed model can reveal more mortality data information and provide a natural quantification of the model parameters uncertainties with no pre-specified constraint required for estimating the cohort effects parameters. The capabilities of the proposed approach are demonstrated through two applications with male and female mortality data. The proposed approach shows remarkable improvements in terms of forecast accuracy compared to the well-known Cairns-Blake-Dowd (CBD) model in the short-, mid- and long-term forecasting using mortality data of several developed countries.

14:30 **Limiting laws for optimal transport plans on finite spaces**
Yoav Zemel

Abstract: Optimal transport is now a popular tool in statistics, machine learning, and data science. The majority of studies regarding the asymptotic properties of optimal transport have focussed on the Wasserstein distance itself, i.e., the optimal objective value. In many situations, however, it is the transport plan (or map) that is more informative, as it allows the practitioner to understand \emph{where}, and not only how much, transport is taking place. We thus study the asymptotics of optimal transport plans in the (practically relevant) case of finite ground space. Possible limiting distributions are derived, and it is shown that the limiting distributions are non-standard with complexity that depends on the degeneracy of the optimal transport linear program and its dual. In particular, if the dual is degenerate (as it is in regular ground spaces), the asymptotic distribution depends on the way the empirical optimal solution is chosen. The results are not specific to optimal transport and hold for general linear programs.

15:00 **Continuous-time multivariate analysis**
Philip Reiss, Biplab Paul

Abstract: The starting point for much of multivariate analysis (MVA) is an n x p data matrix whose n rows represent observations and whose p columns represent variables. Some multivariate data sets, however, may be best conceptualized not as n discrete p-variate observations, but as p curves or functions defined on a common time interval. Such a viewpoint may be useful for multivariate data observed at very high time resolution, with unequal time intervals, and/or with substantial missingness. Here we introduce a framework for extending techniques of multi-

variate analysis to such settings. The proposed framework rests on the assumption that the curves can be represented as linear combinations of basis functions such as B-splines. This is formally identical to the Ramsay-Silverman representation of functional data. But whereas functional data analysis extends MVA methodology to the case of observations that are curves rather than vectors—heuristically, n x p data with p infinite—we are instead concerned with what happens when n is infinite. We demonstrate a new R package that translates the classical MVA methods of principal component analysis, Fisher's linear discriminant analysis, and k-means clustering to the above continuous-time case. The methods are illustrated with a novel perspective on the well-known Canadian weather data set, as well as with an air pollution data set.

Inference for Complex Problem Settings

Organiser: Regina Liu

Chair: Regina Liu

Room: Christian Barnard

13:30 **Query-augmented Active Metric Learning**
Annie Qu

Abstract: We propose an active metric learning method for clustering with pairwise constraints. The proposed method actively queries the label of informative instance pairs, while estimating underlying metrics by incorporating unlabeled instance pairs, which leads to a more accurate and efficient clustering process. In particular, we augment the queried constraints by generating more pairwise labels to provide additional information in learning a metric to enhance clustering performance. Furthermore, we increase the robustness of metric learning by updating the learned metric sequentially and penalizing the irrelevant features adaptively. Specifically, we propose a new active query strategy that evaluates the information gain of instance pairs more accurately by incorporating the neighborhood structure, which improves clustering efficiency without extra labeling cost. In theory, we provide a tighter error bound of the proposed metric learning method utilizing augmented queries compared with methods using existing constraints only. Furthermore, we also investigate the improvement using the active query strategy instead of random selection. Numerical studies on simulation settings and real datasets indicate that the proposed method is especially advantageous when the signal-to-noise ratio between significant features and irrelevant features is low.

14:00 **Additive regression with parametric help**
Hyerim Hong, Young Kyung Lee, *Byeong Uk Park*

Abstract: Additive models have been studied as a way of overcoming theoretical and practical difficulties in estimating a multivariate nonparametric regression function. Several methods have been proposed that ensure the optimal univariate rate one can achieve in estimating univariate nonparametric functions. In this paper a new method is proposed which reduces the constant factor in the first-order approximation of the average squared error of the most successful existing method. The new estimator is based on an orthogonal decomposition of the underlying regression function, with an arbitrarily chosen parametric family, under a special inner product structure arising from the bias formula of the estimator. It is shown that the proposed method entails reduction in the constant factor of the leading bias of the existing method while it retains the same first-order variance. These theoretical findings are confirmed in Monte Carlo experiments.

14:30 **Repro Samples Method for Finite- and Large-Sample Inferences**
Minge Xie

Abstract: This talk presents a novel, general and effective simulation-based approach, called "repro samples method," to conduct statistical inference by creating and studying the performance of artificial samples obtained by mimicking the true observed sample. The artificial samples, referred to as "repro samples," are used to quantify uncertainty and provide confidence sets with guaranteed coverage rates for the target parameter or quantity on a wide range of problems, including many where solutions were previously unavailable or could not be easily obtained. A general framework and supporting theories for both exact and asymptotic inferences are developed. An attractive feature of the development is that it doesn't need to rely on a likelihood or use the large sample central limit theorem, and thus is especially effective for complicated inference problems such as those involving discrete parameters or other problems where the large sample central limit theorem does not apply. The effectiveness of the proposed approach is illustrated through a number of examples, including a case study on a normal mixture model where we construct a finite sample confidence set for the unknown number of components. The performance is also illustrated empirically through simulation and real data examples. Although the development pertains to the settings where the large sample central limit theorem does not apply, it also has direct extensions to the cases where the central limit theorem does hold. (Joint work with Peng Wang)

15:00 **Sufficient Variable Selection via Expected Conditional Hilbert-Schmidt Independence Criterion**
Chenlu (Tracy) Ke

Abstract: We develop a novel model-free variable selection procedure for ultrahigh dimensional data based on a recently proposed independence measure. Compared with sure independence screening methods that only consider marginal dependence between the response and each predictor, our approach inherits the advantages of the new measure and incorporates joint information additionally to achieve sufficient variable selection. As a result, our method is more capable of selecting all the truly active variables, especially those are marginally independent with the response and those involving interactions or nonlinear structures. In addition, the method can handle either continuous or discrete responses with mixed-type predictors. The sure screening property is established under mild conditions, and the superiority of our procedure over existing methods is demonstrated in various simulation studies and an application in real data.

Non-parametric statistical methods for extremes

Organiser: Claudia Kluppelberg

Chair: Claudia Kluppelberg

Room: Leda

13:30 **Neural Networks for Extreme Quantile Regression and Risk Assessment**,
Olivier Pasche, *Sebastian Engelke*

Abstract: Predicting conditional quantiles at extreme probability levels is challenging as classical quantile regression methods tend to fail due to the scarcity of training data in the response's tail region. In the univariate setting, asymptotic results from extreme value theory, such as peaks over threshold, allow to extrapolate quantile values outside the range of the data. We propose to use neural networks for conditional peaks over threshold parameter estimation, yielding a model capable of both extrapolation, and dependence on a large set of predictors. EQRNN models therefore allow modelling extreme conditional quantiles for complex high dimensional mechanisms by taking advantage of the neural network and deep learning methodologies' flexibility and versatility. In particular, they handle both independent and sequentially dependent data. The proposed methodology is motivated by the application to extreme river flows and precipitation, and is shown to outperform existing methods on simulations.

14:00 **High-dimensional extreme quantile regression using partially-interpretable neural networks**
Jordan Richards, *Raphael Huser*

Abstract: Risk management for extreme wildfires requires an understanding of the mechanisms that drive both ignition and spread. Useful metrics for quantifying such risk are extreme quantiles of aggregated burnt area conditioned on predictor variables that describe climate, biosphere and environmental states, as well as the abundance of fuel. Typically these quantiles lie outside the range of observable data and so, for estimation, require specification of parametric extreme value models within a regression framework. Classical approaches in this context utilise linear or additive relationships between predictor and response variables and suffer in either their predictive capabilities or computational efficiency; moreover, their simplicity is unlikely to capture the truly complex structures that lead to the creation of extreme wildfires. In this paper, we propose a new methodological framework for performing extreme quantile regression using artificial neutral networks, which are able to capture complex non- linear relationships and scale well to high-dimensional data. The "black box" nature of neural networks means that they lack the desirable trait of interpretability often favoured by practitioners; thus, we combine aspects of linear, and additive, models with deep learning to create partially interpretable neural networks that can be used for statistical inference but retain high prediction accuracy. To complement this methodology, we further propose a novel point process model for extreme values which overcomes the finite lower-endpoint problem associated with the generalised extreme value class of distributions. Our approach is applied to U.S. wildfire data with a high-dimensional predictor set and we illustrate vast improvements in predictive performance over linear and spline-based regression techniques.

14:30 **Changes in the distribution of observed annual maximum temperatures in Europe**
Ioannis Papastathopoulos, Graeme Auld, Gabriele Hegerl

Abstract: We consider the problem of detecting and quantifying changes in the distribution of the annual maximum daily maximum temperature (TXx) in a large gridded data set of European daily temperature during the years 1950 to 2018. Several statistical models are considered, each of which models TXx using a generalized extreme value (GEV) distribution with the GEV parameters varying smoothly over space. In contrast to several previous studies which fit independent GEV models at the grid box level, our models pull information from neighbouring grid boxes for more efficient parameter estimation. The GEV location and scale parameters are allowed to vary in time using the log of atmospheric CO2 as a covariate. Changes are detected most strongly in the GEV location parameter with the TXx distributions generally shifting towards hotter temperatures. Averaged across our spatial domain, the 100-year return level of TXx based on the 2018 climate is approximately 2°C hotter than that based on the 1950 climate. Moreover, also averaging across our spatial domain, the 100-year return level of TXx based on the 1950 climate corresponds approximately to a 6-year return level in the 2018 climate.

15:00 **Max-linear Bayesian networks**
Claudia Klueppelberg

Abstract: Graphical models can represent multivariate distributions in an intuitive way and, hence, facilitate statistical analysis of high-dimensional data. Such models are usually modular so that high-dimensional distributions can be described and handled by careful combination of lower dimensional factors. Furthermore, graphs are natural data structures for algorithmic treatment. Conditional independence and Markov properties are essential features for graphical models. Moreover, graphical models can allow for causal interpretation, often provided through a recursive system on a directed acyclic graph (DAG) and the max-linear model we introduced in 2018 is a specific example. In this talk I present some conditional independence properties of max-linear Bayesian networks and exemplify the difference to linear networks.

Recent advances in time series

Organiser: Giovanni Motta

Chair: Giovanni Motta

Room: Athena

13:30 High-dimensional dynamic factor models: A selective survey and lines of future research.
Manfred Deistler, Marco Lippi, Brian D.O. Anderson

Abstract: High-Dimensional Dynamic Factor Models are presented in detail: The main assumptions and their motivation, main results, illustrations by means of elementary examples. In particular, the role of singular ARMA models in the theory and applications of High-Dimensional Dynamic Factor Models is discussed. The emphasis of the paper is on model classes and their structure theory, rather than on estimation in the narrow sense. Our aim is not a comprehensive survey. Rather we try to point out promising lines of research and applications that have not yet been sufficiently developed. JEL classification: C50, C55, C53.

14:00 **Semiparametric modeling of multiple quantiles**
Alessandra Luati, Leopoldo Catania

Abstract: We develop a semiparametric model to track a large number of quantiles of a time series. The model satisfies the condition of non crossing quantiles and the defining property of fixed quantiles. A key feature of the specification is that the updating scheme for time varying quantiles at each probability level is based on the gradient of the check loss function, that forms a martingale difference sequence. Theoretical properties of the proposed model are derived, such as weak stationarity of the quantile process and consistency and asymptotic normality of the estimators of the fixed parameters. The model can be applied for filtering and prediction. We also illustrate a number of possible applications such as: i) semiparametric estimation of dynamic moments of the observables, ii) density prediction, and iii) quantile predictions.

14:30 **Asymptotics for Spherical Functional Autoregressions**
Alessia Caponera, *Domenico Marinucci*

Abstract: In this talk, we investigate a class of spherical functional autoregressive processes, and we discuss the estimation of the corresponding autoregressive kernels. In particular, we first establish a consistency result (in sup and mean-square norm), then a quantitative central limit theorem (in Wasserstein distance), and finally a weak convergence result, under more restrictive regularity conditions. We shall also hint at possible extensions, in particular nonparametric testing for stationarity, isotropy and Gaussianity.

15:00 **Sequential on-line detection and classification in 3D Computer Vision**
Olympia Hadjiliadis

Abstract: The topic of interest in this talk is the use of on-line statistical sequential detection techniques in automatic 3D image reconstruction.3 We will begin this presentation by introducing sequential techniques in statistics will stress their importance in applications. In particular, I will contrast the classical hypothesis testing with fixed sample size to sequential decision making and introduce the sequential probability ratio test (SPRT). I will then talk about the problem of quickest detection and introduce the cumulative sum test (CUSUM) and its importance. As an application of the above techniques we will discuss the problem of automatic 3D image reconstruction through laser scan sequential data. We will first apply appropriately tuned CUSUMs to distinguish vertical vs horizontal surfaces. We will then introduce Hidden Markov models to capture vegetation in urban scenes. By applying CUSUMs to detect changes from on Hidden Markov model to another we will be able to identify the beginning of regions of vegetation. By then applying repeated SPRTs, we will be able to identify the ending of these regions. We are thus able to distinguish vertical vs horizontal surfaces as well as regions of vegetation by making use of data sequentially. The second part of this talk is joint work with Dr. Ioannis Stamos.

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited Paper Session 2

Nonparametrics in health related problems

Organiser: Yanyuan Ma
Chair: Yanyuan Ma
Room: Akamas A

16:00 **Multi-institutional Data for Real World Evidence**
Tianxi Cai

Abstract: While clinical trials and cohort studies remain critical sources for studying disease progression and treatment response, they have limitations including the generalizability of the study findings to the real world, the limited ability to examine subgroup effects or test broader hypotheses, and the cost in performing these studies. In recent years, due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with specimen bio-repositories and other research registries, integrated large datasets now open opportunities to generate real world evidence (RWE). Generating reliable RWE with EHR studies, however, remain highly challenging due to heterogeneity across healthcare centers in their patient population and health dynamics. In addition, sharing detailed patient level data cross institutions remains infeasible due to privacy constraints. In this talk, I will discuss federated approaches to study COVID vaccine effects across different patient subgroups to highlight both the value and challenges in using multi-institutional EHR data for RWE.

16:30 **Marginalized frailty-based illness-death model: application to the UK-biobank survival data**
Malka Gorfine

Abstract: The UK Biobank is a large-scale health resource comprising genetic, environmental, and medical information on approximately 500,000 volunteer participants in the United Kingdom, recruited at ages 40–69 during the years 2006–2010. The project monitors the health

and well-being of its participants. This work demonstrates how these data can be used to yield the building blocks for an interpretable risk-prediction model, in a semiparametric fashion, based on known genetic and environmental risk factors of various chronic diseases, such as colorectal cancer. An illness-death model is adopted, which inherently is a semi-competing risks model, since death can censor the disease, but not vice versa. Using a shared-frailty approach to account for the dependence between time to disease diagnosis and time to death, we provide a new illness-death model that assumes Cox models for the marginal hazard functions. The recruitment procedure used in this study introduces delayed entry to the data. An additional challenge arising from the recruitment procedure is that information coming from both prevalent and incident cases must be aggregated. Lastly, we do not observe any deaths prior to the minimal recruitment age, 40. In this work, we provide an estimation procedure for our new illness-death model that overcomes all the above challenges.

17:00 **Estimating Disease Onset from Change Points of Markers Measured with Error**
Tanya Garcia, Unkyung Lee, Raymond Carroll, Yuanjia Wang, Karen Marder

Abstract: Huntington disease is an autosomal dominant, neurodegenerative disease without clearly identified biomarkers for when motor-onset occurs. Current standards to determine motor-onset rely on a clinician's subjective judgment that a patient's extrapyramidal signs are unequivocally associated with Huntington disease. This subjectivity can lead to error which could be overcome using an objective, data-driven metric that determines motor-onset. Recent studies of motor-sign decline --- the longitudinal degeneration of motor-ability in patients --- have revealed that motor-onset is closely related to an inflection point in its longitudinal trajectory. We propose a nonlinear location-shift marker model that captures this motor-sign decline and assesses how its inflection point is linked to other markers of Huntington disease progression. We propose two estimating procedures to estimate this model and its inflection point: one is a parametric method using nonlinear mixed effects model and the other one is a multi-stage nonparametric approach, which we developed. In an empirical study, the parametric approach was sensitive to correct specification of the mean structure of the longitudinal data. In contrast, our multi-stage nonparametric procedure consistently produced unbiased estimates regardless of the true mean structure. Applying our multi-stage nonparametric estimator to PREDICT-HD, a large observational study of Huntington disease, leads to earlier prediction of motor-onset compared to the clinician's subjective judgment.

17:30 **Testing for Heterogeneity in the Utility of a Surrogate Marker**
Layla Parast, Tianxi Cai, Lu Tian

Abstract: In studies that require long-term and/or costly follow-up of participants to evaluate a treatment, there is often interest in identifying and using a surrogate marker to evaluate the treatment effect. While several statistical methods have been proposed to evaluate potential surrogate markers, available methods generally do not account for or address the potential for a surrogate to vary in utility or strength by patient characteristics. Previous work examining surrogate markers has indicated that there may be such heterogeneity i.e., that a surrogate marker may be useful (with respect to capturing the treatment effect on the primary outcome) for some subgroups, but not for others. This heterogeneity is important to understand, particularly if the surrogate is to be used in a future trial to replace the primary outcome. In this paper, we propose a nonparametric approach and estimation procedures to measure the surrogate strength as a function of a baseline covariate W and thus, examine potential heterogeneity in the utility of the surrogate marker with respect to W. Within a potential outcome framework, we quantify the surrogate strength/utility using the proportion of treatment effect on the primary outcome that is explained by the treatment effect on the surrogate. We propose testing procedures to test for evidence of heterogeneity, examine finite sample performance of these methods via simulation, and illustrate the methods using AIDS clinical trial data.

Recent Advances in Time Series Analysis

Organiser: Jens-Peter Kreiss
Chair: Jens-Peter Kreiss
Room: Akamas C

16:00 **Testing equality of spectral density operators for functional linear processes**
Anne Leucht, Efstathios Paparoditis, Theofanis Sapatinas, Daniel Rademacher

Abstract: In this talk, the problem of testing equality of the entire second order structure of two independent functional linear processes is considered. A fully functional L2-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. 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 under the null as well as under the alternative is established. Numerical simulations will be presented showing that, even for small samples, the bootstrap-based test has very good size and power behavior.

16:30 **Statistical methods for the restoration of historical music recordings**
Rainer Dahlhaus

Abstract: In this talk we present methods for the restoration of old shellac recordings from the time period 1927-1950. Different shellacs with the same recording are used and combined to obtain optimal results. The statistical methods used are from robust statistics, nonparametric statistics and time series analysis, and include time warping, noise reduction and filter design. We conclude the talk with some open research problems stimulated by the project.

17:00 **A Bootstrap-Assisted Self-Normalization Approach to Inference in Cointegrating Regressions**
Carsten Jentsch, Karsten Reichold

Abstract: Traditional inference in cointegrating regressions requires tuning parameter choices to estimate a long-run variance parameter. Even in case these choices are "optimal", the tests are severely size distorted. We propose a novel self-normalization approach, which leads to a nuisance parameter free limiting distribution without estimating the long-run variance parameter directly. This makes our self-normalized test tuning parameter free and considerably less prone to size distortions at the cost of only small power losses. In combination with an asymptotically justified vector autoregressive sieve bootstrap to construct critical values, the self-normalization approach shows further improvement in small to medium samples when the level of error serial correlation or regressor endogeneity is large. We illustrate the usefulness of the bootstrap-assisted self-normalized test in empirical applications by analyzing the validity of the Fisher effect in Germany and the United States.

17:30 **Simultaneous Inference for Autocovariances based on Bootstrap Methods**
Alexander Braumann, Jens-Peter Kreiss, Marco Meyer

Abstract: In this talk maximum deviations of sample autocovariances and autocorrelations from their theoretical counterparts over an increasing set of lags are considered. The asymptotic distribution of such statistics for physically dependent stationary time series, which is of Gumbel type, only depends on second order properties of the underlying time series. An obvious question is the asymptotic validity of well-known bootstrap methods such as the autoregressive sieve bootstrap and the multiplicative periodogram bootstrap in this case. We show that this question can be answered positively. However, since speed of convergence of mentioned statistics to its asymptotic distribution is rather slow, higher order properties of the underlying time series, showing up in the asymptotic distribution of the statistic if the set of lags is not increasing, matter for moderate sample sizes. We therefore suggest to use the hybrid periodogram bootstrap (Meyer et al. 2021) for maximum deviations of sample autocovariances and autocorrelations as it correctly captures relevant higher order structures of the underlying time series. Finite sample properties of mentioned bootstrap proposals by simulation are given.

New insights in bandwidth selection

Organiser: Ricardo Cao

Chair: Ricardo Cao

Room: Aphrodite A

16:00 **Bandwidth selection for modal clustering**
Alessandro Casa, *José E. Chacón*, Giovanna Menardi

Abstract: Modal clustering refers to a density-based approach to clustering in which the true population clusters are defined as the domains of attraction of the density modes. Given such a precise population goal, it is possible to define a metric which quantifies the accuracy of its data-based counterpart, naturally suggested as the domains of attraction of a kernel density estimate. Here we present some asymptotic approximations for such a metric, with direct applications to optimal bandwidth selection. The proposed procedures are also investigated in practice via a simulation study.

16:30 **Bandwidth selection for statistical matching and prediction**
Stefan Sperlich, Inés Barbeito, Ricardo Cao

Abstract: We propose a bootstrap method to select a global bandwidth for nonparametric Nadaraya-Watson out-of-sample prediction in impact evaluation, data matching, or scenario simulations. Closed expressions for the bootstrap error criteria are found. This avoids using Monte Carlo approximations, making this selector computationally fast. Asymptotic expressions for the criterion functions and their bootstrap versions are obtained. Both, the derived rate of convergence and our simulation studies show the successful operation of our bandwidth estimator. The method is used to predict nonparametrically the salary of Spanish women if they were paid along the same wage equation as men but conditioned on their own characteristics. An important discrepancy between observed and predicted wages is found exhibiting a serious gender wage gap.

17:00 **Bootstrap bandwidth selection for nonparametric default probability estimation: application to credit risk**
Rebeca Pelaez, Ricardo Cao, Juan Vilar

Abstract: The debts coming from clients with unpaid credits have an important impact in the solvency of banks and other credit institutions. One of the most crucial elements that influence the credit risk is the probability of default (PD). For a fixed time, t, and a horizon time, b, the PD can be defined as the probability that a credit that has been paid until time t, becomes unpaid not later that time t + b. The probability of default conditional on the credit scoring can be written as a transformation of the conditional survival function of the variable "time to default". A nonparametric estimator of the probability of default with double smoothing, both in the covariate and in the time variable, is proposed and studied. It is based on a smoothed version of the Beran conditional survival estimator. Our work focuses on designing an automatic selector of the bivariate bandwidth involved in the smoothed Beran's PD estimator. The issue of confidence regions for the probability of default function PD(t|x) is also addressed. The resampling proposal is an obvious bootstrap method combined with a smoothed bootstrap for the covariate. The bootstrap bandwidth selector is the one that minimises some Monte Carlo approximation of the bootstrap MISE. The bootstrap method is also used to obtain a confidence region that contains the curve PD(t|x) for fixed x and t covering some real interval with probability 1 - alpha. The simulation study carried out over several scenarios concludes that the proposed methods perform satisfactorily. The bandwidth selector is applied to the well-known German Credit data set to estimate and analyse the probability of default conditional on the credit scoring. The confidence region for the function PD(t|x) is also computed to illustrate the results.

17:30 **Bagging cross-validated bandwidth selection in nonparametric regression estimation with application to Big Data**
Daniel Barreiro Ures, *Ricardo Cao*, Mario Francisco Fernández

Abstract: Cross-validation is a well-known and widely used bandwidth selection method in nonparametric regression estimation. However, this technique has two remarkable drawbacks: (i) the large variability of the selected bandwidths, and (ii) the inability to provide results in a reasonable time for very large sample sizes. To overcome these problems, bagging cross-validation bandwidths are analysed in this paper. This approach consists in computing the cross-validation bandwidths for a finite number of subsamples and then rescaling the averaged smoothing parameters to the original sample size. Under a random-design regression model, asymptotic expressions up to a second-order for the bias and variance of the leave-one-out cross-validation bandwidth for the Nadaraya-Watson estimator are obtained. Subsequently, the asymptotic bias and variance and the limit distribution for the bagged cross-validation selector are derived. Suitable choices of the number of subsamples and the subsample size lead to a square root of n rate for the convergence in distribution of the bagging cross-validation selector, outperforming the rate n to the power of -3/10 of leave-one-out cross-validation. Several simulations and an illustration on a real dataset related to the COVID-19 pandemic show the behaviour of our proposal and its better performance, in terms of statistical efficiency and computing time, when compared to leave-one-out cross-validation.

Multivariate and Functional Time series

Organiser: Marco Meyer

Chair: Marco Meyer

Room: Aphrodite B

16:00 **Pivotal tests for relevant differences in the second order dynamics of functional time series**
Anne van Delft, Holger Dette

Abstract: Motivated by the need to statistically quantify differences between modern (complex) data-sets which commonly result as high-resolution measurements of stochastic processes varying over a continuum, we propose novel testing procedures to detect relevant differences between the second order dynamics of two functional time series. In order to take the between-function dynamics into account that characterize this type of functional data, a frequency domain approach is taken. Test statistics are developed to compare differences in the spectral density operators and in the primary modes of variation as encoded in the associated eigenelements. Under mild moment conditions, we show convergence of the underlying statistics to Brownian motions and construct pivotal test statistics. The latter is essential because the nuisance parameters can be unwieldy and their robust estimation infeasible, especially if the two functional time series are dependent. In addition to these novel features, the properties of the tests are robust to any choice of frequency band enabling also to compare energy contents at a single frequency. The finite sample performance of the tests are verified through a simulation study and are illustrated with an application to fMRI data.

16:30 **Testing Linearity for Network Autoregressive Models**
Mirko Armillotta, Konstantinos Fokianos

Abstract: A quasi-score linearity test for continuous and count network autoregressive models is developed. We establish the asymptotic distribution of the test when the network dimension is fixed or increasing, under the null hypothesis of linearity and Pitman's local alternatives. When the parameters are identifiable, the test statistic approximates a chi-square and noncentral chi-square asymptotic distribution, respectively. These results still hold true when the parameters tested belong to the boundary of their space. When we deal with non-identifiable parameters, a suitable test is proposed and its asymptotic distribution is established when the network dimension is fixed. Since, in general, critical values of such test cannot be tabulated, the empirical computation of the p-values is implemented using a feasible bound. Bootstrap approximations are also provided. Moreover, consistency and asymptotic normality of the quasi maximum likelihood estimator is established for continuous and count nonlinear network autoregressions, under standard smoothness conditions. A simulation study and two data examples complement this work.

17:00 **Spectral Inference for Functional Time Series in Hilbert Space**
Daniel Rademacher, Jens-Peter Kreiß, Efstathios Paparoditis

Abstract: A variety of statistics for functional time series allows for a representation as weighted average of corresponding periodogram operators over the frequency domain. We study the consistency and asymptotic distribution of such spectral mean estimates under mild assumptions. We show that the aforementioned asymptotic can be reduced to the asymptotic normality of the sample autocovariance operators, and that the latter condition holds for a large class of weakly dependent functional time series which admit expansions as Bernoulli shifts. The weak dependency is quantified by the condition of L^4-m-approximability and a mixing condition on the 4th order cumulant operators.

17:30 **A frequency domain bootstrap for general multivariate stationary processes**
Marco Meyer, Efstathios Paparoditis

Abstract: Developing valid frequency domain bootstrap procedures for integrated periodogram statistics for multivariate time series is a challenging problem. This is mainly due to the fact that the distribution of such statistics depends on the fourth-order moment structure of the underlying multivariate process in nearly every scenario. Exceptions are some very special cases like nonparametric estimators of the spectral density matrix or Gaussian time series. In contrast to the univariate case, even additional structural assumptions -- such as linearity of the multivariate process or a standardization of the statistic of interest -- do not solve the problem. This paper proposes a new frequency domain bootstrap procedure for multivariate time series, the multivariate frequency domain hybrid bootstrap (MFHB), for integrated periodogram statistics as well as for functions thereof. Asymptotic validity of the MFHB procedure is established for these statistics and for a class of stationary multivariate processes satisfying rather weak dependence conditions ranging clearly beyond linear processes. The finite sample performance of the MFHB is investigated by means of simulations.

Recent advance in covariance estimation and functional data analysis

Organiser: Jane-LingWang

Chair: Derek Young

Room: Christian Barnard

16:00 **Functional Data Analysis of Stochastic Differential Equations**
Victor Panaretos, Neda Mohammadi, Leonardo Santoro

Abstract: We consider the problem of nonparametric estimation of the drift and diffusion coefficients of a Stochastic Differential Equation (SDE), based on n independent replicates on the unit interval, observed irregularly and/or sparsely, and subject to additive noise corruption. We construct estimators by relating the local (drift/diffusion) parameters of the diffusion to their global parameters (mean/covariance, and their derivatives) by means of an apparently novel PDE. This allows us to use methods inspired by functional data analysis, pooling information across the sparsely measured paths Our framework suggests possible further fruitful interactions between FDA and SDE methods in problems with replication.

16:30 **Joint non-parametric estimation of mean and auto-covariances for Gaussian processes**
Tatyana Krivobokova, Paulo Serra

Abstract: Gaussian processes that can be decomposed into a smooth mean function and a stationary autocorrelated noise process are considered and a fully automatic nonparametric method to simultaneous estimation of mean and auto- covariance functions of such processes is developed. Our empirical Bayes approach is data-driven, numerically efficient and allows for the construction of confidence sets for the mean function. Performance is demonstrated in simulations and real data analysis.

17:00 **CoPE sets and Relevant Tubes**
Fabian Telschow, Junting Ren, Armin Schwartzman

Abstract: Recently there has been an increased interest in the statistical analysis of $C(S)$-valued random variables where S is a compact metric space. Different inference methodologies like COnfidence Probability Excursion (COPE) sets, Simultaneous Confidence Bands and statistical tests like relevant difference and bio-equivalence tests have been proposed and successfully applied to functional data. In this talk we demonstrate that the concept of CoPE sets can be used to obtain relevant tube tests for $C(S)$ valued data, explore their theoretical properties and compare them to existing relevance testing methodology.

17:30 **Interpoint-Ranking Sign Covariance for Test of Independence**
Kehui Chen

Abstract: We generalize the sign covariance introduced by Bergsma & Dassios (2014) to multivariate random variables and functional data. The new interpoint-ranking sign covariance is shown to be zero if and only if the two random variables are independent. The test statistic is a U-statistic, whose large sample behavior guarantees that the proposed test is consistent against general types of alternatives. Numerical experiments and data analyses demonstrate the great empirical performance of the proposed method.

Statistical analysis of non-Euclidean and high-dimensional data

Organiser: Byeong Park

Chair: Byeong Park

Room: Leda

16:00 **Total Variation Regularized Frechet Regression for Metric-Space Valued Data**
Zhenhua Lin, Hans-Georg Mueller

Abstract: Non-Euclidean data that are indexed with a scalar predictor such as time are increasingly encountered in data applications, while statistical methodology and theory for such random objects are not well developed yet. To address the need for new methodology in this area, we develop a total variation regularization technique for nonparametric Frechet regression, which refers to a regression setting where a response residing in a generic metric space is paired with a scalar predictor and the target is a conditional Frechet mean. Specifically, we seek to approximate an unknown metric-space valued function by an estimator that minimizes the Frechet version of least squares and at the same time has small total variation, appropriately defined for metric-space valued objects. We show that the resulting estimator is representable by a piece-wise constant function and establish the minimax convergence rate of the proposed estimator for metric data objects that reside in Hadamard spaces. We illustrate the numerical performance of the proposed method for both simulated and real data, including the metric spaces of symmetric positive-definite matrices with the affine-invariant distance and of probability distributions on the real line with the Wasserstein distance.

16:30 **Antipodal Refection Depth (ARD) for Multivariate and Functional Data & Nonparametric Outlier Detection**
Regina Liu

Abstract: Data depth, as a measure of centrality and center-outward ordering, has been developed into a powerful nonparametric alternatives to the classical multivariate and functional data analysis. We introduce a general approach, referred to as antipodal reflection depth (ARD), to refine any existing notion of data depth (referred to as the base depth) to yield a class of new data depth. ARD has these desirable properties: i) It preserves the deepest point and the center-outward ordering along each ray from this deepest point obtained by the base depth; and ii) Its new center-outward ordering captures the relative magnitudes of deviation from all data points to the deepest point. The latter property is generally lacking in the existing notions of data depth due to their location-scale free nature. ARD approach combines the antipodal reflections of the original sample data in the calculation of depth values but draws inferences using only the original data with their associated ARD depth values. This approach is completely data driven and nonparametric. As an immediate application, ARD can be shown to be an effective approach for outlier detection in both multivariate and functional data. Besides simulation studies, ARD is also applied to identify possible anomalous aircraft landings. This is joint work with Dr. Yi Fan, at Amazon Inc.

17:00 **Nonparametric regression on Lie groups with measurement errors**
Jeong Min Jeon, Byeong Uk Park, Ingrid Van Keilegom

Abstract: We study the problem of estimating the regression and density functions for predictors taking values on compact and connected Lie groups and contaminated by measurement errors. Our methodology and theory are based on harmonic analysis on Lie groups. We derive their rates of convergence and asymptotic distributions. We also provide some asymptotic confidence intervals. Our numerical studies show that our estimators outperform the estimators ignoring measurement errors or the geometric structures of Lie groups.

17:30 **Bounds for the asymptotic distribution of the likelihood ratio**
Andreas Anastasiou

Abstract: In this talk, we give an explicit upper bound on the distance to chi-square for the likelihood ratio statistic when the data are realisations of independent and identically distributed random elements. To our knowledge, this is the first explicit bound which is available in the literature. The bound depends on the number of samples as well as on the dimension of the parameter space. We illustrate the bound with various well-known examples, such as samples from an exponential distribution and samples from a normal distribution. This is joint work with Professor Gesine Reinert from the University of Oxford.

19:00 - 20:30 Welcome Reception

TUESDAY 21 JUNE 2022

9:00 - 11:00 Invited Paper Session 3

Semiparametric models

Organiser: Ursula Müller & Valentin Patilea

Chair: Ursula Müller

Room: Akamas A

9:00 **Compositions of discrete random structures in Bayesian nonparametrics**,
Filippo Ascolani, *Antonio Lijoi*, Igor Pruenster, Giovanni Rebaudo

Abstract: Compositions of discrete random probability measures are effective tools in Bayesian nonparametrics for modeling multiple sample data. Hierarchical processes are a noteworthy example, as their infinite-dimensional layers are able to capture latent features that account for data heterogeneity and allow for borrowing of information across different samples. In this talk we consider some general families of hierarchical compositions and will highlight their relevant distributional properties, with a special focus on the induced dependence structure and on the clustering they induce within and across different samples. The presentation will be complemented by some illustrations on simulated and real data.

9:30 **Some recent results on semiparametric transformation models**
Natalie Neumeyer

Abstract: In transformation regression models the response is transformed before fitting a regression model to covariates and transformed response. Some recent results for models with parametric transformation and nonparametric regression are presented. In particular we consider (1) goodness-of-fit tests for the transformation function in nonparametric mean regression, (2) data-based choice of the parametric transformation function in nonparametric regression with one-sided errors, (3) dependence modeling via copula estimation in the case of multivariate responses.

10:00 **Wilks' Theorem for Models Defined by Conditional Moment Equations with Weakly Dependent Data**
Valentin Patilea

Abstract: The aim is the inference for a class of models defined by conditional moment equations, for stationary, strongly mixing series. The empirical likelihood approach is extended to such semiparametric models, which are allowed to include infinite-dimensional nuisance parameters. As an example, the partially linear single-index regression is used for the conditional mean of a one-dimensional series given its past, and the present and past values of a vector of covariates. Moreover, a parametric model for the conditional variance of the series is added to capture further nonlinear effects. We propose an empirical log-likelihood ratio which is allowed to include nonparametric estimators of several functions. It is shown that this ratio behaves asymptotically as if the functions were given, and thus preserves the pivotal property.

10:30 **Data integration in high dimension with multiple quantiles**
Ursula U. Müller

Abstract: This talk deals with the analysis of high dimensional data that come from multiple sources ("experiments") and thus have different possibly correlated responses, but share the same set of predictors. The measurements of the predictors may be different across experiments. We introduce a new regression approach with multiple quantiles to select those predictors that affect any of the responses at any quantile level and to estimate the nonzero parameters. Our approach differs from established methods by being able to handle heterogeneity in data sets as well as heavy-tailed error distributions, two difficulties that are often encountered in complex data scenarios. Our estimator is a minimizer of a penalized objective function, which aggregates the data from the different experiments. We establish model selection consistency and asymptotic normality of the estimator. Simulations and a data application illustrate the advantages of our method in recovering the underlying regression models, which comes from taking the group structure induced by the predictors across experiments and quantile levels into account.
This is joint work with Guorong Dai and Raymond J. Carroll

New trends in high dimensional robust statistics

Organiser: Mohamed Ndaoud

Chair: Mohamed Ndaoud

Room: Akamas C

9:00 **Coding convex bodies under Gaussian noise, and the Wills functional**
Jaouad Mourtada

Abstract: In sequential density estimation, one aims to assign a large probability to a sequence of observations (unknown a priori), close to that of the best a priori distribution within a prescribed class. This prediction problem is intimately connected to that of lossless coding in information theory. A simple integral description of the minimax-optimal error was obtained by Shtarkov in the late 80s. In this work, we study the case of a sequence of real-valued observations, modeled by a subset of the Gaussian sequence model with mean constrained to a general convex body. This can be thought of as an information-theoretic analogue of fixed-design regression. We show that the minimax-optimal error is exactly given by a certain functional of the constraint set from convex geometry called the Wills functional. As a result, we express the optimal error in terms of basic geometric quantities associated to the convex body, namely its intrinsic volumes. After comparing the optimal error to the Gaussian width of the constraint set, we state a fundamental concavity property of the error, and deduce some strong monotonicity properties with respect to noise and sample size.

9:30 **Robust leave-one-out cross-validation for high-dimensional Bayesian models**
Giacomo Zanella

Abstract: Leave-one-out cross-validation (LOO-CV) is a popular method for estimating out-of-sample predictive accuracy. However, computing LOO-CV criteria can be computationally expensive due to the need to fit the model multiple times. In the Bayesian context, importance sampling provides a possible solution but classical approaches can easily produce estimators whose variance is infinite, making them potentially unreliable. Here we propose and analyze a novel mixture estimator to compute Bayesian LOO-CV criteria. Our method retains the simplicity and computational convenience of classical approaches, while guaranteeing finite variance of the resulting estimators. Both theoretical and numerical results are provided to illustrate the improved robustness and efficiency. The computational benefits are particularly significant in high-dimensional problems, allowing to perform Bayesian LOO-CV for a broader range of models. The proposed methodology is easily implementable in standard probabilistic programming software and has a computational cost roughly equivalent to the one of fitting the original model once.

10:00 **All-In-One Robust Estimator of the Gaussian Mean**
Arshak Minasyan

Abstract: We propose a robust-to-outliers estimator of the mean of a multivariate Gaussian distribution that enjoys the following properties: polynomial computational complexity, high breakdown point, minimax rate optimality (up to logarithmic factor) and asymptotical efficiency. Non-asymptotic risk bound for the expected error of the proposed estimator is dimension-free and involves only the effective rank of the covariance matrix. Moreover, we show that the obtained results can be extended to sub-Gaussian distributions, as well as to the cases of unknown rate of contamination or unknown covariance matrix.

10:30 **Outlier detection in networks**
Olga Klopp, Geneviève Robin, Solenne Gaucher

Abstract: Outliers arise in networks due to different reasons such as fraudulent behavior of malicious users or default in measurement instruments and can significantly impair network analyses. In addition, real-life networks are likely to be incompletely observed, with missing links due to individual non-response or machine failures. Therefore, identifying outliers in the presence of missing links is an important problem in network analysis. In this work we introduce a new algorithm to detect outliers in a network and simultaneously predict the missing links. The proposed method is statistically sound: under fairly general assumptions, this algorithm exactly detects the outliers, and achieves the best known error for the prediction of missing links with polynomial computational cost.

Recent advances in change point and time series problems in high dimensions

Organiser: Hira Koul & Zhou Zhou

Chair: Liudas Giraitis

Room: Aphrodite A

9:00 **High-dimensional changepoint estimation with heterogeneous missingness**
Tengyao Wang, Bertille Follain, Richard Samworth

Abstract: We propose a new method for changepoint estimation in partially-observed, high-dimensional time series that undergo a simultaneous change in mean in a sparse subset of coordinates. Our first methodological contribution is to introduce a `MissCUSUM' transformation (a generalisation of the popular Cumulative Sum statistics), that captures the interaction between the signal strength and the level of missingness in each coordinate. In order to borrow strength across the coordinates, we propose to project these MissCUSUM statistics along a direction found as the solution to a penalised optimisation problem tailored to the specific sparsity structure. The changepoint can then be estimated as the location of the peak of the absolute value of the projected univariate series. In a model that allows different missingness probabilities in different component series, we identify that the key interaction between the missingness and the signal is a weighted sum of squares of the signal change in each coordinate, with weights given by the observation probabilities. More specifically, we prove that the angle between the estimated and oracle projection directions, as well as the changepoint location error, are controlled with high probability by the sum of two terms, both involving this weighted sum of squares, and representing the error incurred due to noise and the error due to missingness respectively. A lower bound confirms that our changepoint estimator, which we call MissInspect, is optimal up to a logarithmic factor. The striking effectiveness of the MissInspect methodology is further demonstrated both on simulated data, and on an oceanographic data set covering the Neogene period.

9:30 **Multiple change-points detection in generalized linear models**
Yuehua Wu

Abstract: In this talk, we focus on the problem of multiple change points estimation in GLMs in which both number of change points and their locations are unknown. We propose a simultaneous multiple change points estimation method which first partitions the data sequence into several segments to construct a new design matrix, secondly convert the multiple change points estimation problem into a variable selection problem, and then estimate the regression coefficients by maximizing a penalized likelihood function. The consistency of the coefficient estimator is established in which the number of coefficients can diverge as the sample size goes to infinity. The nonzero coefficient estimates provide the information about which segments potentially contain a change point. An algorithm is provided to estimate the multiple change points. Simulation studies are conducted for both logistic and log-linear models. A real data application is also presented.

10:00 **Empirical process theory for stochastic processes using the functional dependence measure**
Stefan Richter

10:30 **Time-Varying Instrumental Variable Estimation**
Liudas Giraitis, George Kapetanios, Massimiliano Marcellino

Abstract: We develop non-parametric instrumental variable estimation and inferential theory for econometric models with possibly endogenous regressors whose coefficients can vary over time either deterministically or stochastically, and the time-varying and uniform versions of the standard Hausman exogeneity test. After deriving the asymptotic properties of the proposed procedures, we assess their finite sample performance by means of a set of Monte Carlo experiments, and illustrate their application by means of an empirical example on the Phillips curve

Recent contributions in statistical analysis of nonstationary processes

Organiser: Anna Dudek

Chair: Anna Dudek

Room: Aphrodite B

9:00 **Some contributions to harmonizable time series analysis**
Jean-Marc Freyermuth

Abstract: Harmonizable time series are natural extensions of stationary time series with a spectral decomposition whose components are correlated. Thus the covariance function of an harmonizable time series is bivariate and admits a two-dimensional Fourier decomposition (Loève spectrum). They form a broad class of non-stationary processes that has been subject of investigation for a long time starting with Loève (1948-1963), Rozanov (1959) and Cramèr (1961). In this talk, we will introduce harmonizable VARMA time series and discuss how to generate them from given Loève spectra. Then, we will present some recent theoretical results for nonparametric spectral analysis based on replicated realizations of spatiotemporal processes that are locally time-harmonizable. An illustration based on EEG recordings arising from an experiment in neuropsychology will be provided.

09:30 **Spectral Non-linear Granger Causality**
Hernando Ombao

Abstract: One of the key goals in analyzing multivariate time series is to understand the nature of dependence between its components. This is key in understanding how the entire brain network is engaged during cognitive processing and how this network may be disrupted due to some shock to the system such as a stroke or an epileptic seizure. In this talk, we propose a novel algorithm to extract frequency-band specific and non-linear Granger causality (Spectral NLGC) connections in multivariate time series. The advantages of our model over traditionally used VAR based models is shown using simulations over synthetic non-linear time series at different noise levels. We applied the Spectral NLGC method to uncover potential non-linear dynamics in an epileptic seizure EEG data set recorded from a patient diagnosed with left temporal lobe epilepsy. Spectral NLGC gives new meaningful insights into frequency specific connectivity changes at the onset of epileptic seizure. Results of both simulated and real life time series confirms the viability of the proposed algorithm as a good tool for exploring directed frequency-specific connectivity.

10:00 **Spectral density estimation for nonstationary data with nonzero mean function**
Anna Dudek, Lukasz Lenart

Abstract: We introduce a new approach for nonparametric spectral density estimation based on the subsampling technique, which we apply to the important class of nonstationary time series. These are almost periodically correlated sequences. In contrary to existing methods our technique does not require demeaning of the data. On the simulated data examples we compare our estimator of spectral density function with the classical one. Additionally, we propose a modified estimator, which allows to reduce the leakage effect.

Some recent advances in estimation and inference for complex data models

Organiser: Moulinath Banerjee

Chair: Moulinath Banerjee

Room: Christian Barnard

9:00 **Unlinked monotone regression**
Charles Doss, Fadoua Balabdaoui, *Cécile Durot*

Abstract: We consider so-called univariate unlinked monotone regression. In an unlinked regression model, we observe predictors and response variables, but we do not see the actual pairing of predictors and responses; the predictors and responses may be gathered entirely independently. This problem may seem to be a hopeless one. However, if we assume that the regression function is monotone (say, increasing), and furthermore assume that we know the distribution of the regression error terms, then we can derive a nonparametric regression estimator. We develop an algorithm for its computation. We show that this estimator converges to the true regression function, and give the estimator's rate of convergence in L_1 loss. The rate of convergence depends on the smoothness of the error distribution. We discuss extensions to the case in which the distribution of the errors is unknown. We demonstrate our estimator's use on synthetic data and on data from the US Consumer Expenditure Survey.

9:30 **A General Modeling Framework for Network Autoregressive Processes**
George Michailidis

Abstract: A general flexible framework for Network Autoregressive Processes (NAR) is developed, wherein the response of each node in the network linearly depends on its past values, a prespecified linear combination of neighboring nodes and a set of node-specific covariates. The corresponding coefficients are node-specific, and the framework can accommodate heavier than Gaussian errors with spatial-autoregressive, factor based or in certain settings general covariance structures. We provide a sufficient condition that ensures the stability (stationarity) of the underlying NAR that is significantly weaker than its counterparts in previous work in the literature. Further, we develop ordinary and (estimated) generalized least squares estimators for both fixed, as well as diverging number of network nodes, and also provide their ridge regularized counterparts that exhibit better performance in large network settings, together with their asymptotic distributions. We derive their asymptoic distributions that can be used for testing various hypotheses of interest to practitioners. We also address the issue of misspecifying the network connectivity and its impact on the aforementioned asymptotic distributions of the various NAR parameter estimators. The framework is illustrated on both synthetic and real air pollution data.

10:00 **Fast Network Community Detection with Profile-Pseudo Likelihood Methods**
Ji Zhu

Abstract: he stochastic block model is one of the most studied network models for community detection. It is known that most algorithms proposed for fitting the stochastic block model likelihood function cannot scale to large-scale networks. One prominent work that overcomes this computational challenge is Amini et al. (2013), which proposed a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have a convergence guarantee. In this talk, we present a novel likelihood based approach that decouples row and column labels in the likelihood function, which enables a fast alternating maximization; the new method is computationally efficient and has provable convergence guarantee. We also show that the proposed method provides strongly consistent estimates of the communities in a stochastic block model. As demonstrated in simulation studies, the proposed method outperforms the pseudo-likelihood approach in terms of both estimation accuracy and computation efficiency, especially for large sparse networks. We further consider extensions of the proposed method to handle networks with degree heterogeneity and bipartite properties. This is joint work with Jiangzhou Wang, Jingfei Zhang, Binghui Liu, and Jianhua Guo.

Computer-intensive methods and dependent data

Organiser: Dimitris Politis

Chair: Dimitris Politis

Room: Leda

9:00 **Estimation and inference via the integrated copula spectrum**
Holger Dette, *Stanislav Volgushev*, Tobias Kley, Marc Hallin, Yuichi Goto, Ria Van Hecke

Abstract: Frequency domain methods form a ubiquitous part of the statistical toolbox for time series analysis. In recent years, considerable interest has been given to the development of new spectral methodology and tools capturing dynamics in the entire joint distributions and thus avoiding the limitations of classical, L2-based spectral methods. Most of the spectral concepts proposed in that literature suffer from one major drawback, though: their estimation requires the choice of a smoothing parameter, which has a considerable impact on estimation quality and poses challenges for statistical inference. In this paper, associated with the concept of copula-based spectrum, we introduce the notion of copula spectral distribution function or integrated copula spectrum. This integrated copula spectrum retains the advantages of copula-based spectra but can be estimated without the need for smoothing parameters. We discuss such estimators, along with a thorough theoretical analysis, based on a functional central limit theorem, of their asymptotic properties. We show how one can leverage these results to test various hypotheses that cannot be addressed by classical spectral methods, such as the lack of time-reversibility or asymmetry in tail dynamics.

9:30 **Bootstrap for Dynamical Systems**
*Kasun Fern*ando, *Nan Zou*

Abstract After its establishment in the late 19th century through the efforts of Poincaré and Lyapunov, the theory of dynamical systems was applied to study processes in the real world. Despite their deterministic nature, dynamical systems can still exhibit incomprehensibly chaotic behaviors and seemingly random patterns. Consequently, a dynamical system is usually represented by a probabilistic model of which the unknown parameters must be estimated using statistical methods. To measure the uncertainty of such parameter estimation, this talk will develop the bootstrap for dynamical systems and establish its consistency and second-order efficiency via continuous Edgeworth expansions.

10:00 **High-dimensional Change-point Detection Using Generalized Homogeneity Metrics**
Xianyang Zhang, Shubhadeep Chakraborty

Abstract: Change-point detection has been a classical problem in statistics, finding applications in various fields. A nonparametric change-point detection procedure is concerned with detecting abrupt distributional changes in the data generating distribution, rather than only mean changes. We consider the problem of detecting an unknown number of change-points in an independent sequence of high-dimensional observations and testing for the significance of the estimated change-point locations. Our approach rests on nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point location estimator via defining a cumulative sum process in an embedded Hilbert space. As the main theoretical innovation, we rigorously derive its limiting distribution under the high-dimension medium sample size framework. Subsequently, we combine our statistics with the idea of wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to several other existing procedures is illustrated via both simulated and real datasets.

10:30 **Rate of convergence in regenerative bootstrap**
Patrice Bertail, François Portier

Abstract: We recall the principles of the approximate regenerative bootstrap for positive recurrent Markov chains . The rate of convergence of approximate regenerative bootstrap distribution is linked to the rate of convergence of the transition density for the uniform L2 norm over a well chosen small set. We show how it is possible to obtain exponential inequality which allow to obtain explicit rate of convergence.

Recent advances in semiparametric inference

Organiser: Christophe Ley

Chair: Olivier Thas

Room: Athena

9:00 **Experiences with the super learner**
Thomas Alexander Gerds

Abstract A super learner is stacked regression in disguise, a meta machine learning algorithm that is used for semi-parametric inference. The refrain of the super learner song goes ``Stack strong learners together, the result performs asymptotically as well as the best possible weighted combination.'' Behind this is the cross-validation selector which has nice asymptotic properties. In this talk, I will discuss some experiences with the super learner, here in particular, the role of proper scoring rules, ``garbage in, garbage out'', applications in right censored survival analysis, and the leave-one-out bootstrap to reduce the Monte-Carlo error of the level-one data.

09:30 **Using Laguerre Polynomials in Semiparametric Estimation with an Application to Quantile Regression and Epidemics**
Alexander Kreiss, Ingrid Van Keilegom

Abstract: In this talk we will investigate the potential of Laguerre polynomials for two estimation tasks: Quantile regression and estimation of the basic reproduction number of a disease. In quantile regression we model the conditional quantiles of a quantity of interest as a linear function of observed covariates plus noise. We suppose that the conditional distribution of the noise conditional on the covariates can be adequately approximated by Laguerre polynomials. This allows for a rich class of densities. In our estimation procedure we estimate the quantile parameter and the noise density simultaneously. In the talk we will investigate if and under which conditions this extra flexibility of the estimator can lead to efficiency gains. In our second application, in epidemics, quantities like the reproduction number depend on the incubation period (time from infection to symptom onset) and/or the generation time (time until a new person is infected from another infected person). Similarly to above we approximate the densities of these quantities through a formulation using Laguerre polynomials. This results in a simple semi-parametric sieve-estimation method for estimation of these distributions. We provide detailed theory for consistency and illustrate the finite sample performance for small datasets via a simulation study. This is joint work with Ingrid Van Keilegom (KU Leuven).

10:00 **Semi-parametric quantile regression**
Anneleen Verhasselt, Ewnetu Worku, Irène Gijbels

Abstract: Quantile regression allows to assess the effects of covariates, not only on a location parameter (such as a mean or median) but also on specific percentiles of the conditional distribution. In recent years, a large family of flexible two-piece asymmetric distributions where the location parameter coincides with a specific quantile of the distribution has been studied. We propose a semiparametric procedure to estimate the conditional quantile curves of two-piece asymmetric distributions. We use a local likelihood estimation technique, via which the effect of a covariate on the location, scale, and index of the conditional distribution can be assessed.

10:30 **Semiparametric Methods for Covariate Adjustment for GPC Effect Sizes**
Olivier Thas

Abstract: Generalised pairwise comparisons (GPC) is gaining more and more attention as an effect size in clinical trials. It can take several forms (e.g net benefit, win ratio, win odds, probabilistic index) and can be defined for a single outcome as well as for multiple outcomes. The estimation of this effect size, and its properties, is still an ongoing research area, and correcting the GPC effect size for covariates is considered important for furhter exteding the application range of GPCs, but it has not yet attracted much attention. We have developed flexible semiparametric methods for analysing the net benefit with adjustment for baseline covariates. These methods are based on Probabilistic Index Models and they are easy to implement. In this talk, we will outline the construction of the methods and demonstrate them on a case study.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk: Richard A. Davis

Chair: Dimitris Politis
Room: Akamas A

Statistical Learning of Multivariate Extremes

Abstract: A spectral clustering algorithm for analyzing the dependence structure of multivariate extremes is proposed. More specifically, we focus on the asymptotic dependence of multivariate extremes characterized by the angular or spectral measure in the multivariate regular variation setting. Our work studies the theoretical performance of spectral clustering based on a random k-nearest neighbor graph constructed from an extremal sample, i.e., the angular part of random vectors for which the radius exceeds a large threshold. In particular, we derive the asymptotic distribution of extremes arising from a linear factor model and prove that, under certain conditions, spectral clustering can consistently identify the clusters of extremes arising in this model. Leveraging this result we propose a simple consistent estimation strategy for learning the angular measure. Our theoretical findings are complemented with numerical experiments illustrating the finite sample performance of our methods. (This is joint work with Marco Avella Medina and Gennady Samorodnitsky.)

12:30 - 13:30 Lunch Break

13:30 - 15:30 Invited Paper Session 4

Recent advances in change point analysis and high-dimensional time series
Organiser: George Michailidis
Chair: George Michailidis
Room: Akamas A

13:30 **Estimation of High-Dimensional Markov-Switching VAR Models with an Approximate EM Algorithm**
Ali Shojaie, Xiudi Li, Abolfazl Safikhani

Abstract: Regime shifts in high-dimensional time series arise naturally in many applications, from neuroimaging to finance. This problem has received considerable attention in low-dimensional settings, with both Bayesian and frequentist methods used extensively for parameter estimation. The EM algorithm is a particularly popular strategy for parameter estimation in low-dimensional settings, although the statistical properties of the resulting estimates have not been well understood. Furthermore, its extension to high-dimensional time series has proved challenging. To overcome these challenges, we propose an approximate EM algorithm for Markov-switching VAR models that leads to efficient computation and also facilitates the investigation of asymptotic properties of the resulting parameter estimates. We establish the consistency of the proposed EM algorithm and investigate its performance via simulation studies.

14:00 **Two-sample tests for relevant differences in the eigenfunctions of covariance operators**
Alexander Aue, Holger Dette, Gregory Rice

Abstract: This talk deals with two-sample tests for functional time series data, which have become widely available in conjunction with the advent of modern complex observation systems. Here, particular interest is in evaluating whether two sets of functional time series observations share the shape of their primary modes of variation as encoded by the eigenfunctions of the respective covariance operators. To this end, a novel testing approach is introduced that connects with, and extends, existing literature in two main ways. First, tests are set up in the relevant testing framework, where interest is not in testing an exact null hypothesis but rather in detecting deviations deemed sufficiently relevant, with relevance determined by the practitioner and perhaps guided by domain experts. Second, the proposed test statistics rely on a self-normalization principle that helps to avoid the notoriously difficult task of estimating the long-run covariance structure of the underlying functional time series. The main theoretical result of this paper is the derivation of the large-sample behavior of the proposed test statistics. Empirical evidence, indicating that the proposed procedures work well in finite samples and compare favorably with competing methods, is provided through a simulation study, and an application to annual temperature data.

14:30 **Frequency-domain graphical models for multivariate time series**
Sumanta Basu

Abstract: Graphical models offer a powerful framework to capture intertemporal and contemporaneous relationships among the components of a multivariate time series. For stationary time series, these relationships are encoded in the multivariate spectral density matrix and its inverse. We will present adaptive thresholding and penalization methods for estimation of these objects under suitable sparsity assumptions. We will discuss new optimization algorithms and investigate consistency of estimation under a double-asymptotic regime where the dimension of the time series increases with sample size. If time permits, we will introduce a frequency-domain graphical modeling framework for multivariate nonstationary time series that captures a new property called conditional stationarity.

15:00 **Robust change point and change plane estimation in fixed and growing dimensions under heavy tailed errors**
Moulinath Banerjee

Abstract: We present a number of new findings about the canonical change point estimation problem. We first study the estimation of a change point on the real line in a simple stump model using the robust Huber estimating function which interpolates between the ell_1 (absolute deviation) and ell_2 (least squares) based criteria. While the ell_2 criterion has been studied extensively, its robust counterparts and in particular, the ell_1 minimization problem have not. We derive the limit distribution of the estimated change point under the Huber estimating function and compare it to that under the ell_2 criterion. Theoretical and empirical studies indicate that it is more profitable to use the Huber estimating function under heavy tailed errors as it leads to smaller asymptotic confidence intervals at the usual levels compared to the ell_2 criterion. We also compare the ell_1 and ell_2 approaches in a parallel setting, where one has m independent single change point problems and the goal is to control the maximal deviation of the estimated change points from the true values, and establish that the ell_1 estimation criterion provides a superior rate of convergence to the ell_2, and that this relative advantage is driven by the heaviness of the tail of the error distribution. Finally, we derive minimax optimal rates for the change plane estimation problem in growing dimensions and demonstrate that Huber estimation attains the optimal rate while the ell_2 scheme produces a rate sub-optimal estimator for heavy tailed errors. In the process of deriving our results, we establish a number of properties about the minimizers of compound Binomial and compound Poisson processes which are of independent interest. This is joint work with Debarghya Mukherjee and Ya'acov Ritov.

Recent innovations in Bayesian Nonparametrics

Organiser: Subhashis Ghoshal
Chair: Subhashis Ghoshal
Room: Akamas C

13:30 **Bayesian mode estimation and rate acceleration through a two-stage procedure**
William Weimin Yoo

14:00 **Characterization and Bayesian estimation of sparse regression parameters under linear inequality constraints**
Anindya Roy

Abstract: Modern statistical problems often involve linear inequality constraints on model parameters. Ignoring natural parameter constraints usually results in less efficient statistical procedures. A large number of parameters along with restrictions can significantly increase model complexity. Such high complexity can only be tackled by making simplifying assumptions about the parameter. To this end, we define a notion of 'sparsity' for parameters restricted to closed convex polyhedral cones described by linear inequalities. We allow our framework to be flexible so that the number of restrictions may be much higher than the number of parameters. Such situations arise in function estimation under shape restriction. We show that the proposed notion of sparsity agrees with the usual notion of sparsity in the unrestricted case and prove the validity of the proposed definition as a measure of sparsity. The proposed sparsity measure also allows us to generalize popular priors for sparse vector estimation to the constrained case. We illustrate the usefulness of the proposed prior through examples.

14:30 **Bayesian sensitivity analysis for a missing outcomes model**
Stephanie van der Pas, Bart Eggen, Aad van der Vaart

Abstract: (Joint work with Bart Eggen and Aad van der Vaart) When outcome data is missing, even in randomized controlled trials, drawing causal conclusions is not straightforward. Solutions usually rest on unverifiable assumptions, thereby creating a new problem. Sensitivity analysis allows us to assess the robustness of study conclusions to these assumptions. We study a model where outcomes are missing due to participants dropping out post-intervention and adopt a Bayesian approach to incorporate prior beliefs on selection bias parameters. We provide theoretical guarantees for the eventual estimate of the mean outcome. We show two Bernstein-von Mises theorems for different parametrizations of the model, using a Dirichlet process prior and a normalized extended gamma process prior respectively.

Nonparametric approaches in Econometrics

Organiser: Enno Mammen
Chair: Enno Mammen
Room: Aphrodite A

13:30 **Flexible Covariate Adjustments in Regression Discontinuity Designs**
Claudia Noack, Christoph Rothe, Tomasz Olma

Abstract: Empirical regression discontinuity (RD) studies often use covariates to increase the precision of their estimates. In this paper, we propose a novel class of estimators that use such covariate information more efficiently than the linear adjustment estimators that are currently used widely in practice. Our approach can accommodate a possibly large number of either discrete or continuous covariates.

It involves running a standard RD analysis with an appropriately modified outcome variable, which takes the form of the difference between the original outcome and a function of the covariates. We characterize the function that leads to the estimator with the smallest asymptotic variance, and show how it can be estimated via modern machine learning, nonparametric regression, or classical parametric methods. The resulting estimator is easy to implement because tuning parameters can be chosen as in a conventional RD analysis. An extensive simulation study illustrates the performance of our approach.

14:00 **Multiplicative Deconvolution In a Bivariate Stochastic Volatility Model**
Sergio Brenner Miguel

Abstract: In this talk, we use discrete time observations of a bivariate diffusion process to construct a nonparametric estimator of the density of the underlying, unobserved bivariate volatility process. Based on the estimation of the Mellin transform of the volatility density and a spectral cut-off regularisation of the inverse Mellin transform, we propose a fully data-driven density estimator where the anisotropic choice of the spectral cut-off parameter is dealt by a model selection approach. Furthermore, we study the risk of our estimator and show its adaptivity up to a negligible term. We demonstrate the reasonable performance of our estimator using a Monte-Carlo simulation and consider several examples for the volatility processes.

14:30 **Square-root-2 estimation for smooth eigenvectors of matrix-valued functions**
Giovanni Motta, Wei Biao Wu, Mohsen Pourahmadi

Abstract: Modern statistical methods for multivariate time series rely on the eigendecomposition of time-varying covariance as well as the spectral density matrices. The curse of indeterminacy (or miss-identification) of smooth eigenvector functions has not received much attention. We resolve this important problem and recover smooth trajectories by examining the distance between successive eigenvectors. We change the sign of the next eigenvector if its distance with the current one is larger than the square root of 2. In the case of distinct eigenvalues, this simple method delivers smooth eigenvectors. In the case of coalescing (intersecting) eigenvalues, additional swapping and bridging around the coalescing points are needed. We establish consistency and rates of convergence for the proposed smooth eigenvector estimators. We provide simulation results and applications to real data, where we show that our approach is needed to obtain smooth eigenvectors.

15:00 **Estimation of Group Structures in Individual Fixed Effects Models**
Enno Mammen, Ralf Wilke, Kristina Zapp

Abstract: In this paper we consider a fixed effect linear panel model where N individuals are observed over a time period of T time points. The model contains covariates X_{it} that depend on time t and individual i and covariates Z_i that depend only on the individual. Furthermore, the model contains an additional additive unobserved fixed effect v_i. Thus the model writes as Y_{it} = X_{it} beta + Z_i gamma + v_i + epsilon_{it}. Without further assumptions, in this model the parameter gamma is not identified. We make the following assumption for identification: for a non-vanishing unknown fraction of the individuals the fixed effects belong to a finite cluster set. For the remaining individuals the fixed effect v_i is allowed to be different for each individual. We show that under weak additional assumptions in this setting the parameter gamma can be estimated with parametric rate square root of (NT). We will apply our model to labor market data. The talk reports on joint work with Ralf Wilke (Copenhagen) and Kristina Zapp (Mannheim).

Resampling methods in non-standard situations

Organiser: Patrice Bertail

Chair: Patrice Bertail

Room: Aphrodite B

13:30 **Parameters on the boundary in predictive regression**
Iliyan Georgiev, Giuseppe Cavaliere

Abstract: We consider bootstrap inference in predictive (or Granger-causality) regressions when the parameter of interest is on the boundary of the parameter space, here defined by means of a smooth inequality constraint. For instance, this situation occurs when the definition of the paramer space formalizes the dichotomy of either no predictability or sign-restricted predictability, and the null hypothesis of no predictability (resp., no Granger causality) is tested against the one-sided alternative contained in the parameter space . We show that in this context constrained estimation gives rise to bootstrap statistics whose limit distribution is, in general, random, and thus distinct from the limit null distribution of the original statistics of interest. This is due to both (i) the location of a true parameter value on the boundary of the parameter space, and (ii) the possible non-stationarity of the posited predicting (resp. Granger-causing) variable. We discuss a modification of the standard fixed-regressor wild bootstrap scheme where the bootstrap parameter space is shifted by a data-dependent function, thus allowing us to eliminate the boundary as a source of limiting bootstrap randomness. With possible non-stationarity as the only remaining source of limiting randomness, we prove validity of the associated bootstrap inference in the cases where the posited predicting variable is either I(1) or I(0). Our approach, which is initially presented in a simple location model, has bearing on inference in parameter-on-the boundary situations beyond the predictive regression.

14:00 **General M-Estimator Processes and their m out of n Bootstrap with Functional Nuisance Parameters.**
Anouar Abdeldjaoued Ferfache

Abstract: In the present talk, we consider the problem of the estimation of a parameter theta, in Banach spaces, maximizing some non-smooth criterion function which depends on an unknown nuisance parameter h, possibly infinite-dimensional. The classical estimation methods are mainly based on maximizing the corresponding empirical criterion by substituting the nuisance parameter with some nonparametric estimator. We show that the M-estimators converge weakly to maximizers of Gaussian processes under rather general conditions. The conventional bootstrap method fails, in general, to consistently estimate the limit law. We show that the m out of n bootstrap, in this extended setting, is weakly consistent under conditions similar to those required for weak convergence of the M-estimators. The aim of this work is therefore to extend the existing theory on the bootstrap of the M-estimators. Examples of applications from the literature are given to illustrate the generality and usefulness of our results. Finally, we investigate the performance of the methodology for small samples through a short simulation study.

14:30 **Bootstrap on Null-recurrent Markov Chains**
Carlos Fernández

Abstract: Two bootstrap methods, previously presented for positive recurrent Markov chains, are introduced for the null recurrent case. A variation of the Central Limit Theorem for a random number of summands is also presented.

15:00 **Model-free Bootstrap and Conformal Prediction in Regression**
Dimitris Politis, Yiren Wang

Abstract: Predictive inference under a general regression setting is gaining more interest in the big-data era. In terms of going beyond point prediction to develop prediction intervals, two main threads of development are conformal prediction and Model-free prediction. Recently, Chernozhukov, Wuthrich and Zhu (2021) proposed a conformal prediction approach exploiting the uniformization procedure inherent in the Model-free Bootstrap of Politis (2015). It is of interest to compare and further investigate the performance of the two methods. In the paper at hand, we contrast the two approaches via theoretical analysis and numerical experiments with a focus on conditional coverage of prediction intervals. We discuss suitable scenarios for applying each algorithm, underscore the importance of conditional vs. unconditional coverage, and show that, under mild conditions, the Model-free bootstrap yields prediction intervals with guaranteed better conditional coverage compared to naive quantile estimation. We also extend the concept of `pertinence' of prediction intervals of Politis (2015) to the nonparametric regression setting, and give concrete examples where its importance emerges under finite sample scenarios. Finally, we define the new notion of `conjecture testing' that is the analog of hypothesis testing as applied to the prediction problem; we devise a modified conformal score to allow conformal prediction to handle one-sided `conjecture tests', and compare to the Model-free bootstrap.

Advances in functional data analysis

Organiser: Siegfried Hörmann

Chair: Siegfried Hörmann

Room: Christian Barnard

13:30 **Relative perturbation bounds with applications to empirical covariance operators**
Johannes Moritz Jirak

Abstract: The goal of this paper is to establish relative perturbation bounds, tailored for empirical covariance operators. Our main results are expansions for empirical eigenvalues and spectral projectors, leading to concentration inequalities and limit theorems. One of the key ingredients is a specific separation measure for population eigenvalues, which we call the relative rank, giving rise to a sharp invariance principle in terms of limit theorems, concentration inequalities and inconsistency results. Our framework is very general, requiring only p>4 moments and allows for a huge variety of dependence structures.

14:00 **Fast and Fair Simultaneous Confidence Bands for Functional Parameters**
Dominik Liebl

Abstract: Quantifying uncertainty using confidence regions is a central goal of statistical inference. Despite this, methodologies for confidence bands in Functional Data Analysis are still underdeveloped compared to estimation and hypothesis testing. In this work, we present a new methodology for constructing simultaneous confidence bands for functional parameter estimates. Our bands possess a number of striking qualities: (1) they have a nearly closed-form expression and thus are fast to compute, (2) they can be constructed adaptively according to a desired criterion, where we focus on the fairness constraint of false positive rate balance across partitions of the bands' domain which facilitates both global and local interpretations, and (3) they do not require an estimate of the full covariance function and thus can be used in the case of fragmentary functional data. Simulations show the excellent finite-sample behavior of our bands in comparison to existing alternatives. The practical use of our bands is demonstrated in two case studies on sports biomechanics and fragmentary growth curves.

14:30 **Estimation of Functional ARMA Models**
Thomas Kuenzer

Abstract: Functional auto-regressive moving average (FARMA or ARMAH) models allow for flexible and natural modelling of functional time series. While there are many results on pure autoregressive (FAR) models in Hilbert spaces, results on estimation and prediction of FARMA models are considerably more scarce. We devise a simple two-step method to estimate ARMA models in separable Hilbert spaces. Estimation is based on dimension-reduction using principal components analysis of the functional time series. We establish consistency of the proposed estimators under simple assumptions by employing a data-driven criterion to select the dimensionality of the principal component subspaces used in the estimation procedure. The empirical performance of the estimation algorithm is evaluated in a simulation study, where it performs better than competing methods.

15:00 **On optimal prediction of missing functional data with memory**
Germain Van Bever, Lauri Viitasaari, Pauliina Ilmonen, Tommi Sottinen, Nourhan Shafik

Abstract: In this talk, we consider the problem of reconstructing missing parts of functions based on their observed segments. It provides, for Gaussian processes and arbitrary bijective transformations thereof, theoretical expressions for the L2-optimal reconstruction of the missing parts. These functions are obtained as solutions of explicit integral equations. In the discrete case, approximations of the solutions provide consistent expressions of all missing values of the processes. In the case of Gaussian processes with a parametric covariance structure, the estimation can be conducted separately for each function, and yields nonlinear solutions in presence of memory. We explore the empirical performances of the discrete approximation in simulated and real examples.

New frontiers in network data analysis

Organiser: Marianna Pensky & Srijan Sengupta

Chair: Marianna Pensky

Room: Leda

13:30 **Partially-Exchangeable Multilayer Stochastic Block Models**
Daniele Durante, Francesco Gaffi, Antonio Lijoi, Igor Prünster

Abstract: There is an increasing availability of complex network data encoding connectivity information among a set of nodes, often belonging to different layers. For example, in bill co-sponsorship networks, the nodes correspond to political candidates, whereas layers denote the party of affiliation, and the goal is to infer grouping structures made by candidates with similar behaviors in co-sponsoring bills. Although it is reasonable to expect that such blocks would potentially overlap with the partition defined by layers (e.g., party of affiliation), this assumption is often too strong and fails to learn sub-blocks within each layer as well as grouping structures that span across multiple layers. To incorporate these mixed architectures while accounting for layer information in a principled manner, I will present a new class of partially-exchangeable multilayer stochastic block models which relies on a hierarchical random partition prior for the node allocations to groups driven by the urn scheme of a hierarchical normalized completely random measure (H-NRMI) or a hierarchical Pitman-Yor process (H-PYP). The partial exchangeability assumption among nodes according to layer partitions allows to infer both within-and across-layer blocks, while preserving probabilistic coherence, principled uncertainty quantification and formal inclusion of prior information from layer membership. The mathematical tractability of such priors further allows to analytically derive and compare predictive within- and across-layer co-clustering probabilities, thereby providing conditions on hyperparameters to enforce interpretable grouping structures coherent with the observed multilayer network. The applied potentials of this new class of Bayesian nonparametric models are illustrated in political and criminal network studies.

14:00 **Population-level Balance in Signed Networks**
Ji Zhu

Abstract: Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for signed networks have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. In this paper, we introduce a statistically principled latent space approach for modeling signed networks and accommodating the well-known balance theory, i.e., "the enemy of my enemy is my friend" and "the friend of my friend is my friend". The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates, which are further verified through simulation studies. We also apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II.

14:30 **A nonparametric test of co-spectrality in networks**
Srijan Sengupta

Abstract: We live in an interconnected world where network-valued data arise in many domains, and, fittingly, statistical network analysis has emerged as an active area in the literature. However, the topic of hypothesis testing in networks has received relatively less attention. In this work, we consider the problem where one is given two networks and the goal is to test whether the given networks are cospectral, i.e., they have the same non-zero eigenvalues. Cospectral graphs have been well studied in graph theory and computer

science. Cospectrality is relevant in real-world networks since it implies that the two networks share several important path-based properties, such as the same number of closed walks of any given length, the same epidemic threshold, etc. However, to the extent of our knowledge, there has not been any formal statistical inference work on this topic. We propose a non-parametric test of co-spectrality by leveraging some recent developments in random matrix theory. We develop two versions of the test — one based on an asymptotic bound and one based on bootstrap resampling. We establish theoretical results for the proposed test and demonstrate its empirical accuracy using synthetic networks sampled from a wide variety of models as well as several well-known real-world network datasets. This work is in collaboration with Chetkar Jha (University of Pennsylvania) and Indrajit Jana (Indian Institute of Technology, Bhuvaneshwar).

15:00 **Discovering underlying dynamics in time series of networks**
Avanti Athreya, Zachary Lubberts, Youngser Park, Carey Priebe

Abstract: Understanding dramatic changes in the evolution of networks is central to statistical network inference, the importance of which has been underscored by the challenges of predicting and distinguishing pandemic-induced transformations in organizational and communication networks. We address this problem of multi-sample network inference with a model in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits low-dimensional manifold structure under a suitable notion of distance. This distance can be approximated by a measure of separation between the networks themselves, and there exist consistent Euclidean representations for the underlying network structure, as characterized by these distances, at any given time. These Euclidean representations permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classically familiar setting.
We illustrate our methodology with real and synthetic data, and identify change points corresponding to massive shifts in pandemic policy in a communication network of a large organization.

Density and regression estimation under non standard conditions

Organiser: Valentin Patilea

Chair: Valentin Patilea

Room: Athena

13:30 **Empirical Risk Minimization under Random Censorship**
Stephan Clemencon

Abstract: We consider the classic supervised learning problem where a continuous non-negative random label Y (a random duration) is to be predicted based upon observing a random vector X by means of a regression rule with minimum least square error. In various applications, ranging from industrial quality control to public health through credit risk analysis for instance, training observations can be right censored, meaning that, rather than on independent copies of (X,Y), statistical learning relies on a collection of n independent realizations of the triplet (X, min{Y, C}, d), where C is a nonnegative random variable with unknown distribution, modeling censoring and d indicates whether the duration is right censored or not. As ignoring censoring in the risk computation may clearly lead to a severe underestimation of the target duration and jeopardize prediction, we consider a plug-in estimate of the true risk based on a Kaplan-Meier estimator of the conditional survival function of the censoring C given X, referred to as Beran risk, in order to perform empirical risk minimization. It is established, under mild conditions, that the learning rate of minimizers of this biased/weighted empirical risk functional is of the same order as that which can be attained in absence of censoring. Beyond theoretical results, numerical experiments are presented in order to illustrate the relevance of the approach developed.

14:00 **Partly Linear Instrumental Variables Regression Without Smoothing on the Instruments**
Elia Lapenta, Jean-Pierre Florens

Abstract: We propose a new estimation method for partly linear models identified by Instrumental Variables (IVs). The estimation is based on a class of Generically Comprehensively Revealing functions. Compared to methods available in the literature that smooth on the IVs, our estimation does not smooth on the instruments and thus requires the selection of less tuning parameters. Furthermore, it does not suffer from a curse of dimensionality on the IVs. We show that our procedure is equivalent to a classical estimation method that smooths on the IVs but keeps the bandwidth fixed as the sample size increases. We obtain convergence rates for the estimator of the nonparametric part of the model and the asymptotic normality of the estimator of the parametric components. To deal with the ill-posedness of the inverse problem, we use a Landweber-Friedman regularization. This is a simple iterative method that does not require the inversion of a large matrix whose dimension increases with the sample size. We finally study the implementation of our procedure and propose a data-driven selection of the regularization and the smoothing parameters.

14:30 **Improved estimation of semiparametric cure regression models via presmoothing**
Eni Musta, Ingrid Van Keilegom, Valentin Patilea

Abstract: We consider survival data with a cure fraction, which means that some subjects will never experience the event of interest. Accounting for the possibility of cure has recently become highly relevant in oncology as progress is being made for treatment of different cancer types. However, cure rate estimation is statistically challenging in the presence of censored time-to-event data. A common model in this context is the mixture cure model which consists of two submodels: one for the probability of being cured and one for the survival of the uncured subjects, conditional on a set of covariates. In particular, the logistic-Cox model is a frequent choice. Because of the latent cure status, maximum likelihood estimation is performed by means of the iterative EM algorithm and suffers from several prob-

lems when the sample size is limited. We propose an alternative estimation method that is based on presmoothing by first constructing a nonparametric estimator and then projecting it in the desired parametric class, e.g. logistic. Once the cure probabilities are estimated, one can estimate the survival of the uncured subjects in a second step. We investigate the advantages and theoretical properties of this new approach and show through simulations that, in the logistic-Cox model, it outperforms the maximum likelihood estimator for small and moderate sample sizes. The method is used to analyse melanoma survival data.

15:00 **Semiparametric efficiency under identifiability constraints**
Mélanie Zetlaoui

Abstract: This talk considers the problem of parameter estimation in semiparametric models under identifiable constraints. In parametric models it is possible to compute Fisher information by a correct reparametrization of the original (non-identifiable) parameter. The purpose of the talk is to propose a general framework for computing information bounds in the semiparametric case. This problem appears in many semiparametric models. We present here the case of single index models (and their generalization), nonparametric latent variable models and mixture models.

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited Paper Session 5

New nonparametric and semiparametric methods on learning data with complex structures

Organiser: Wen Zhou

Chair: Chao Zheng

Room: Akamas A

16:00 **Confidence sets for Causal Discovery**
Mladen Kolar

Abstract: Causal discovery procedures are popular methods for discovering causal structure across the physical and social sciences. However, most procedures for causal discovery take in data, and output an estimated causal model. In this paper, we propose a procedure for forming confidence sets for problems in causal discovery. Specifically, we propose a procedure which returns a set of causal orderings which are not ruled out by the data. When the true generative procedure falls within a certain class, we show that asymptotically the true ordering will be contained in the returned set with some user specified probability.
Joint work with Sam Wang and Mathias Drton

16:30 **Data Integration Via Analysis of Subspaces (DIVAS)**
Jan Hannig

Abstract: A major challenge in the age of Big Data is the integration of disparate data types into a data analysis. That is tackled here in the context of data blocks measured on a common set of experimental subjects. This data structure motivates the simultaneous exploration of the joint and individual variation within each data block. This is done here in a way that scales well to large data sets (with blocks of wildly disparate size), using principal angle analysis, careful formulation of the underlying linear algebra, and differing outputs depending on the analytical goals. Ideas are illustrated using cancer and neuroimaging data sets. Joint work with J.S. Marron, Jack Prothero, and Meilei Jiang. There is an earlier paper on this subject that will be partially discussed here: https://doi.org/10.1016/j.jmva.2018.03.008

17:00 **Generalized dynamic factor models for spatio-temporal random fields on a network**
Chao Zheng

Abstract: High dimensional datasets containing records of spatio-temporal structures are of interest in many application areas—e.g., brain imaging, meteorology, marketing research. In this work, we consider a dimensionality reduction technique using the generalised dynamic factors models. Differently from the extant approaches (available in the time series setting), we have to take into account the spatial dependence. To this end, we define our generalised factor model by working on the spectral theory of random fields on Z^3. Unlike the time series setting, where time domain representation theory (using the Wold's theorem and the concept of innovations) are derived, unfortunately, the same argument cannot apply to the spatio-temporal random fields as there is no unique definition of the concept "innovation". In this paper, we established the rigorous definitions of multivariate space time processes, its spectral representation theory, and a consistent estimator of the spectral density. Numerical and real-data examples are also provided.

Expanding Statistical Frontiers with Nonparametric Methods

Organiser: Tanya Garcia

Chair: Tanya Garcia

Room: Akamas C

16:00 **Nonparametric instrumental regression with right censored duration outcomes**
Ingrid Van Keilegom, Jad Beyhum, Jean-Pierre Florens

Abstract: This paper analyzes the effect of a discrete treatment Z on a duration T. The treatment is not randomly assigned. The confounding issue is treated using a discrete instrumental variable explaining the treatment, and independent of the error term of the model. Our framework is nonparametric and allows for random right censoring. This specification generates a nonlinear inverse problem and the average treatment effect is derived from its solution. We provide local and global identification properties that rely on a nonlinear system of equations. We propose an estimation procedure to solve this system and derive rates of convergence and conditions under which the estimator is asymptotically normal. When censoring makes identification fail, we develop partial identification results. Our estimators exhibit good finite sample properties in simulations. We also apply our methodology to the Illinois Reemployment Bonus Experiment.

16:30 **Robust and Efficient Estimation under Nonignorable Missing Response**
Yanyuan Ma

Abstract: We consider the estimation problem in a regression setting where the outcome variable is subject to nonignorable missingness and identifiability is ensured by the shadow variable approach. We propose a versatile estimation procedure where modeling of missingness mechanism is completely bypassed. We show that our estimator is easy to implement and we derive the asymptotic theory of the proposed estimator. We also investigate some alternative estimators under different scenarios. Comprehensive simulation studies are conducted to demonstrate the finite sample performance of the method. We apply the estimator to a children's mental health study to illustrate its usefulness.

17:00 **Covariance and phase recovery for multivariate time series**
Irène Gannaz, Sophie Achard

Abstract: Many applications have to deal with multivariate time series. In neuroscience, recordings of brain activity consist in time series associated to brain regions. The measures of long-memory properties and the coupling between time series have shed lights for understanding brain mechanisms. We propose an analytic wavelets based procedure which recovers jointly long-memory properties, the modulus of long-run covariance between time series, and phases. The procedure can be applied in particular to estimate the parameters of the multivariate fractional Brownian motion. We illustrate the procedure on a real fMRI dataset.

17:30 **Bayesian variable selection based on empirical likelihood for ultra-high dimensional data**
Xinlei Wang, Can Xu, Yichen Cheng

Abstract: A great amount of literature has shown that the development of variable selection techniques can enable an efficient and interpretable analysis of high dimensional data, which are ubiquitous nowadays. However, variable selection involving ultra-high dimensional data, where the number of covariates p is (much) large than the sample size n, remains a highly challenging task. Furthermore, many popular methods based on linear regression models assume Gaussian random noise. In the semi-parametric domain, under the ultra-high dimensional setting, we propose a Bayesian empirical likelihood method for variable selection, which requires no distributional assumptions but only estimating equations. Motivated by Chang et al. (2018, Annals of Statistics) on doubly penalized empirical likelihood (EL), we introduce priors to regularize both regression parameters and Lagrange multipliers associated with the estimating equations, to promote sparse learning. We further develop an efficient Markov chain Monte Carlo sampling algorithm based on the active set idea, which has been proved to be useful in reducing computational burden in several existing studies. The proposed method not only inherits merits from both Bayesian and EL inferences but also has superior performance in both the prediction and variable selection, as shown in our numerical studies.

New perspectives on nonparametric estimation of intensity and density functions
Organiser: Maria Dolores Martinez-Miranda
Chair: Maria Dolores Martinez-Miranda
Room: Aphrodite A

16:00 **Statistical learning for general point processes and applications to intensity and density estimation**
Christophe A.N. Biscio, Ottmar Cronie, Mehdi Moradi

Abstract: Point processes generalise iid random samples by allowing a random sample size and/or the sample points to be dependent. In this talk, we present the first statistical learning framework for general point processes and show how to use it for intensity, and thereby density, estimation under these non-iid circumstances. Our new approach is based on a subtle combination of two new concepts in point process theory: prediction errors and cross-validation. The general idea is to split a point process in two, through thinning, and estimate parameters by predicting one part using the other. This allows us to introduce a variety of loss functions not only suitable for standard spatial statistical problems but for general estimation settings, without imposing the iid assumptions. We will introduce our framework for a general statistical audience, not necessarily familiar with point process theory, and illustrate how our methods can be used for intensity and density estimation. In particular, we will show numerically that it substantially outperforms state of the art in bandwidth selection for kernel intensity estimators. If time permits, we will also indicate how our new methodology could be applied in other point process settings.

16:30 **The modal age of Statistics**
José E. Chacón

Abstract: As a statistical parameter, the mode has been traditionally considered little relevant, at least as compared to its more prominent cousins, the mean and the median. However, a number of statistical problems have recently found an unexpected solution by inspecting them through a "modal perspective". These include classical tasks such as clustering or regression. This has led to a renewed interest in estimation and inference for the mode. In this talk we will survey the standard approaches to mode estimation and explore the consequences of applying this modern modal methodology to other, seemingly unrelated, fields.

17:00 **Kernel estimation methods beyond planar point processes**
Maria Isabel Borrajo García

Abstract: Point processes are a branch of spatial statistics focused on the analysis of stochastic processes generating patterns of events, which are random in number and location. One of the main areas of interest within point processes are planar point processes, where we assume the support is a subregion of the Euclidean plane, with all the advantages in terms of metric, derivative definition, … that it implies. In this context, where those bivariate locations are the given information, different estimates of the intensity function, which measures the expected number of events per unit measure, have been proposed. However, there exist a wider world beyond them, for instance, covariate information may be used to improve our estimates or other different support domains may be of interest for various applications. In this talk we present different kernel intensity estimates for those alternative scenarios: we detail a kernel intensity estimator using covariates, as well as other kernel intensity estimates for point processes defined on different supports, such as linear networks or the d-dimensional sphere.

17:30 **Superefficient estimation of future conditional hazards based on marker information**
Alex Isakson, Jens Nielsen, Enno Mammen, Cécile Proust-Lima

Abstract: We introduce a new concept for forecasting future events based on marker information. The model is based on a nonparametric approach with counting processes featuring so-called high-quality markers. Despite the model having nonparametric parts, we show that we attain a parametric rate of uniform consistency and uniform asymptotic normality. In usual nonparametric scenarios, reaching such a fast convergence rate is not possible, so one can say that our approach is superefficient. This theory is then used to construct simultaneous confidence bands directly for the hazard rate. In this talk, we will validate the methodology by comparing it with a joint modelling approach in a simulation study, and illustrate its use for the computation of individual dynamic predictions in the context of primary biliary cirrhosis (PBC) of the liver.

Financial Econometrics

Organiser: Genaro Sucarrat

Chair: Genaro Sucarrat

Room: Aphrodite B

16:00 **Inference on multiplicative component GARCH without any small-order moment**
Christian Francq, Baye Matar Kandji, Jean-Michel Zakoian

Abstract: In multiplicative component GARCH models, the volatility is decomposed into the product of two factors which often received interpretations in terms of "short run" (high frequency) and "long run" (low frequency) components. While two-component volatility models are widely used in applied works, some of their theoretical properties remain unexplored. We show that the strictly stationary solutions of such models do not admit any small-order finite moment, contrary to classical GARCH. It is shown that the strong consistency and the asymptotic normality of the Quasi-Maximum Likelihood estimator hold despite the absence of moments. Tests for the presence of a long-run volatility relying on the asymptotic theory and a bootstrap procedure are proposed. Our results are illustrated via Monte Carlo experiments and real financial data.

16:30 **Testing hypotheses on the innovations distribution in semi-parametric conditional volatility models**
Jean-Michel Zakoian, Christian Francq

Abstract: The paper considers the problem of testing assumptions on the innovations in GARCH-type models. We propose tests of different hypotheses: adequacy of a parametric quantile, mean-median equality, symmetry of extreme quantiles and zero-median in presence of a conditional mean. The tests rely on the asymptotic distribution of the empirical distribution function of the residuals. They are generally model-free (though not estimation-free) and thus are simple to implement. Efficiency comparisons are made and a numerical study based on simulated and real data is provided.

17:00 **Subgeometrically ergodic autoregressions with autoregressive conditional heteroskedasticity**
Mika Meitz, Pentti Saikkonen

Abstract: In this paper, we consider subgeometric ergodicity of univariate nonlinear autoregressions with autoregressive conditional heteroskedasticity (ARCH). The notion of subgeometric ergodicity was introduced in the Markov chain literature in 1980s and it means that the transition probability measures converge to the stationary measure at a rate slower than geometric; this rate is also closely related

to the convergence rate of beta-mixing coefficients. While the existing literature on subgeometrically ergodic autoregressions assumes a homoskedastic error term, this paper provides an extension to the case of conditionally heteroskedastic ARCH-type errors, considerably widening the scope of potential applications. Specifically, we consider suitably defined higher-order nonlinear autoregressions with possibly nonlinear ARCH errors and show that they are, under appropriate conditions, subgeometrically ergodic at a polynomial rate. An empirical example using energy sector volatility index data illustrates the use of subgeometrically ergodic AR–ARCH models.

17:30 **Robust estimation and inference for time-varying unconditional volatility,**
Rickard Sandberg, Genaro Sucarrat

Abstract: The unconditional volatility of financial return is often time-varying. To model this, a common approach is to decompose the volatility multiplicatively into a non-stochastic process, and a de-volatilitised stochastic process. We prove the consistency and asymptotic (CAN) normality of the single-step Quasi Maximum Likelihood Estimator (QMLE) for the parameters of a broad class of specifications of the non-stochastic process. Next, we derive a simple but robust and consistent estimator of the coefficient covariance. The exact specification of the stochastic process (often assumed to be a GARCH model) need not be estimated or known, and it can even be non-stationary in the distribution. This is important and useful in empirical applications, since financial returns are frequently characterised by a non-stationary zero-process. Additionally, our results can be used in multi-step estimation procedures in which the parameters of stochastic process, e.g. a GARCH model, are estimated in a second step. Due to the assumptions we rely upon, our results extend directly to the Multiplicative Error Model (MEM) interpretation of volatility models. In other words, our results can also be applied to the modelling of the time-varying unconditional mean of non-negative processes like volume, duration, realised volatility and unemployment

Nonparametric estimation

Organiser: Alexander Goldenshluger

Chair: Alexander Goldenshluger

Room: Christian Barnard

16:00 **Semi-parametric Bernstein-von Mises theorem for densities with jump under a mixture prior**
Natalia Bochkina, Judith Rousseau, J.-B. Salomond, Johan van der Molen Moris

Abstract: We consider the problem of density estimation where the density has unknown lower support point, under local asymptotic exponentiality, from a Bayesian perspective. We state general sufficient conditions for the local concentration of the marginal posterior of the lower support (Bernstein - von Mises type theorem) which has a faster 1/n rate and exponential distribution with a random shift, under an adaptive estimation of the unknown density. We constructed an adaptive mixture prior for a decreasing density with the following properties: a) posterior distribution of the density with known lower support point concentrates at the minimax rate, up to a log factor, b) the density is estimated consistently, uniformly in a neighbourhood of the lower support point, c) marginal posterior distribution of the lower support point of the density has shifted exponential distribution in the limit. Consistent estimation of the unknown density at the lower support point is important, as it is the scale parameter of the limiting shifted exponential distribution.
In particular, to ensure that the density is asymptotically consistent pointwise in a neighbourhood of the lower support point, instead of a usual Dirichlet mixture weights, we consider a non-homogeneous Completely Random Measure mixture. The general conditions for the BvM type result we have are different from those by Knapik and Kleijn (2013); the latter don't hold for a hierarchical mixture prior we consider. We illustrate performance of this approach on simulated data, and apply it to model distribution of bids in procurement auctions.

16:30 **Statistical guarantees for high dimensional generative models Arnak Dalalyan**,
Victor-Emmanuel Brunel, Nicolas Schreuder

Abstract: We introduce a convenient framework for studying (adversarial) generative models from a statistical perspective. It consists in modeling the generative device as a smooth transformation of the unit hypercube of a dimension that is much smaller than that of the ambient space and measuring the quality of the generative model by means of an integral probability metric. In the particular case of an integral probability metric defined through a smoothness class, a risk bound is established, quantifying the role of various parameters. In particular, it clearly shows the impact of dimension reduction on the error of the generative model.

17:00 **Semiparametric estimation of McKean-Vlasov SDEs**
Denis Belomestny, Vytaute Pilipauskaite, Mark Podolskij

Abstract: In this talk we study the problem of semiparametric estimation for a class of McKean-Vlasov stochastic differential equations. Our aim is to estimate the drift coefficient of a MV-SDE based on observations of the corresponding particle system. We propose a semiparametric estimation procedure and derive the rates of convergence for the resulting estimator. We further prove that the obtained rates are essentially optimal in the minimax sense.

17:30 **Semiparametric ordered inference for conditional distributions**
Ori Davidov

Abstract: In a variety of applications researchers are interested in comparing or ordering two or more treatment groups conditional on some auxiliary covariates. To address this large and important class of problems in a flexible way we develop efficient semiparametric constrained estimation and testing procedures for ordering conditional distributions. More specifically, a new class of semiparametric regression models is introduced and studied. Necessary and sufficient conditions for the ordering of distributions, by the likelihood ratio order, with and without

covariates are derived. The parameters of conditionally ordered distributions can be estimated using semi-infinite programming and an algorithm for doing so is proposed and proven to converge. A consistent likelihood ratio test for, or against, an ordering is proposed. The operating characteristics of the new methodology are explored using simulations and illustrated using a data example.

Statistical methodology for complex and heterogeneous data

Organiser: Alexander Aue

Chair: Alexander Aue

Room: Leda

16:00 **Graphical models for nonstationary time series**
Suhasini Subbarao, Sumanta Basu

Abstract: We propose NonStGM, a general nonparametric graphical modeling framework for studying dynamic associations among the components of a nonstationary multivariate time series. It builds on the framework of Gaussian Graphical Models (GGM) and stationary time series Graphical models (StGM), and complements existing works on parametric graphical models based on change point vector autoregressions (VAR). Analogous to StGM, the proposed framework captures conditional noncorrelations
(both intertemporal and contemporaneous) in the form of an undirected graph. In addition, to describe the more nuanced nonstationary relationships among the components of the time series, we introduce the new notion of conditional nonstationarity/stationarity and incorporate it within the graph. This can be used to search for small subnetworks that serve as the ``source'' of nonstationarity in a large system. We explicitly connect conditional noncorrelation and stationarity between and within components of the multivariate time series to zero and Toeplitz embeddings of an infinite-dimensional inverse covariance operator. In the Fourier domain, conditional stationarity and noncorrelation relationships in the inverse covariance operator are encoded with a specific sparsity structure of its integral kernel operator. We show that these sparsity patterns can be recovered from finite-length time series by node-wise regression of discrete Fourier Transforms (DFT) across different Fourier frequencies. We demonstrate the feasibility of learning NonStGM structure from data using simulation studies.

16:30 **Preprocessing functional data by a factor model approach**
Siegfried Hörmann, Fatima Jammoul

Abstract: We consider functional data which are measured on a discrete set of observation points. Often such data are measured with noise, and then the target is to recover the underlying signal. Commonly this is done with some smoothing approach, e.g. kernel smoothing or spline fitting. While such methods act function by function, we argue that it is more accurate to take into account the entire sample for the data preprocessing. To this end we propose to fit factor models to the raw data. We show that the common component of the factor model corresponds to the sig- nal which we are interested in, whereas the idiosyncratic component is the noise. Under mild technical assumptions we demonstrate that our estimation scheme is uniformly consistent. Moreover, we provide a tailor-made testing framework for the assumption of iid noise. From a theoretical standpoint our approach is elegant, because it is not based on smoothness assumptions and generally permits a realistic framework. The practical implementation is easy because we can resort to existing tools for factor models. Our empirical investigations provide convincing results.

17:00 **Bootstrapping Spectral Statistics in High Dimensions**
Miles Lopes, Andrew Blandino, Alexander Aue

Abstract: Statistics derived from the eigenvalues of sample covariance matrices are called spectral statistics, and they play a central role in multivariate testing. Although bootstrap methods are an established approach to approximating the laws of spectral statistics in low-dimensional problems, such methods are relatively unexplored in the high-dimensional setting. The aim of this work is to focus on linear spectral statistics as a class of prototypes for developing a new bootstrap in high dimensions, a method we refer to as the spectral bootstrap. In essence, the proposed method originates from the parametric bootstrap and is motivated by the fact that in high dimensions it is difficult to obtain a non-parametric approximation to the full data-generating distribution. From a practical standpoint, the method is easy to use and allows the user to circumvent the difficulties of complex asymptotic formulas for linear spectral statistics. In addition to proving the consistency of the proposed method, we present encouraging empirical results in a variety of settings. Lastly, and perhaps most interestingly, we show through simulations that the method can be applied successfully to statistics outside the class of linear spectral statistics, such as the largest sample eigenvalue and others.

17:30 **Functional Sequential Treatment Allocation**
David Preinerstorfer, Anders Bredahl Kock, Bezirgen Veliyev

Abstract: The multi-armed bandit literature has mainly focused on targeting the arm with the highest mean. In this talk I discuss minimax expected regret optimality results for situations where the observational structure is that of a multi-armed bandit problem, but where instead of the arm with the highest mean one targets the arm with the highest value of a functional of the cdf of the outcome distribution, e.g., a quantile, trimmed mean, or a welfare measure.

TUESDAY 21 JUNE 2022

Exotic Testing and Inverse Problems

Organiser: Robert Mnatsakanov

Chair: Robert Mnatsakanov

Room: Athena

16:00 **What is the resolution of a microscope?**
Frank Werner

Abstract: As a general rule of thumb the resolution of a light microscope (i.e. the ability to discern objects) is predominantly described by the full width at half maximum (FWHM) of its point spread function (psf)---the diameter of the blurring density at half of its maximum. Classical wave optics suggests a linear relationship between FWHM and resolution also manifested in the well known Abbe and Rayleigh criteria, dating back to the end of 19th century. However, during the last two decades conventional light microscopy has undergone a shift from microscopic scales to nanoscales. This increase in resolution comes with the need to incorporate the random nature of observations (light photons) and challenges the classical view of discernability, as we argue in this paper. Instead, we suggest a statistical description of resolution obtained from such random data. Our notion of discernability is based on statistical testing whether one or two objects with the same total intensity are present. For Poisson measurements we get linear dependence of the (minimax) detection boundary on the FWHM, whereas for a homogeneous Gaussian model the dependence of resolution is nonlinear. Hence, at small physical scales modeling by homogeneous gaussians is inadequate, although often implicitly assumed in many reconstruction algorithms. In contrast, the Poisson model and its variance stabilized Gaussian approximation seem to provide a statistically sound description of resolution at the nanoscale. Our theory is also applicable to other imaging setups, such as telescopes.

16:30 **Goodness of Fit Testing for Point Processes in Survival Analysis**
Umut Can, Roger Laeven, Estate Khmaladze

Abstract: Suppose we are given an observed path from a temporal point process, and we would like to test whether a particular parametric model for the conditional intensity of this process matches the observed path. We propose a novel approach for conducting such goodness of fit tests. The idea is to consider the compensated point process, where the compensator is estimated parametrically, and to transform this process into a Poisson process compensated by its own estimated compensator. Then it is sufficient to know the asymptotic behavior of the latter process in order to test the goodness of fit of the former, for a wide class of parametric intensity models. We demonstrate the applicability of our approach through Monte Carlo simulations of Aalen-type survival processes, with and without censoring.

17:00 **Identification of convex polygons and non-negative probability measures on convex polygons from moment data**
Farhad Jafari, Robert Mnatsakanov, Satwik Pani

Abstract: Let K be a bounded closed convex polygon in R2 and μ be a nonnegative probability measure supported on K. If p is a polynomial of degree d on K, then μ has an associated weighted moment multi-sequence a_p. V. Tchakaloff has proved that there exist N = N (2, d) points x_1,...,x_N in K and nonnegative constants b_j , j = 1,...,N such that a_p is the b-weighted sum of p(x_j). That is, there are N points in K such that μ - integrals of polynomials of up to total degree d can be written as a linear combination of Dirac measures supported at the points x_j. Since any closed convex polygon is the closed convex hull of its extreme points (vertices), the points x_j are convex linear combinations of the vertices. Hence the above representation can be rephrased in terms of the polynomials evaluated at the vertices. Coefficients b_j and convexity coefficients c_ij can be expressed in terms of moments and vice versa (the opposite direction is trivial). Linear independence gives an upper bound on how many moments would be necessary to fully identify the extreme points of these polygons. In this talk, we relate the moments of the indicator function of polygons to moments of measures supported on its vertices, and use exact relations to identify polygonal regions from moments supported on its vertices. Using these relationships some patterns of moments of nonnegative probability measures supported on polygons are identified. An example of these results are applied to tomography data are presented.

17:30 **A unified method of estimation in some statistical inverse problems**
Robert Mnatsakanov

Abstract: In many statistical inverse problems, like deconvolution, demixing, multiplicative-censoring models, as well as in problems associated with recovering the images in the Computed Tomography, the Laplace transform of unobserved distribution (function) of actual interest can be easily derived (estimated) from the values of corresponding Laplace transform of the given data-set. In this talk we introduce a unified method that could be applied in several statistical inverse problems. For instance, in the problems of recovering a multivariate function from the finite number of values of its Radon transforms, we introduce two novel approximations and estimators of an unknown function. Furthermore, using the empirical counterparts of the Laplace transform of underlying function a new estimate of the Radon transform itself is derived. Under smoothed conditions on unknown function, the uniform convergence of the proposed constructions are established. Several other applications of proposed approximations in the information theory and statistics will be discussed as well. The graphical illustrations are provided to demonstrate the accuracy of the approximates.

18:15 - 19:15 General Meeting
Room: Akamas A

WEDNESDAY 22 JUNE 2022

9:00 - 10:00 Keynote Talk: Aurore Delaigle
Chair: Marianna Pensky
Room: Akamas A

Estimation of the Distribution of Episodically Consumed Foods Measured with Error

Abstract: Dietary data collected from 24-hour dietary recalls are observed with significant measurement errors, in the sense that they are imprecise measurements of the long term intake of nutrients. In the nonparametric curve estimation literature, a lot of effort has been devoted to designing methods that are consistent under contamination by noise. However, some foods such as alcohol or fruits are consumed only episodically, and may not be consumed during the day when the 24-hour recall is administered. If the food is consumed on data collection day, the reported intake behaves like a contaminated version of a latent variable related to usual intake; if the food is not consumed on data collection day, the reported intake is equal to zero. For example, in the 2011--2013 Australian Health Survey, more than 5\% of the reported intakes of folic acid, caffeine and alcohol were equal to zero. Such data can be represented by a two part model, for which parametric techniques are well established, such as the National Cancer Institute approach. However, existing nonparametric errors-in-variables methods cannot deal with the excess zeros present in the data. We present new estimators of the distribution of such episodically consumed food data.

10:00 - 11:00 Contributed Paper Session 2

Censored and missing data

Chair: Amichai Painsky
Room: Akamas A

10:00 Nonparametric survival estimation with missing not at random censoring indicators
Mikael Escobar-Bach, Olivier Goudet

Abstract: In the presence of right-censored data with random covariates, the conditional Kaplan-Meier estimator (also referred to as the Beran estimator) consistently estimates the conditional survival function. However, it relies on the knowledge of each individual censoring status, which might be missing in practice. In this talk, we thus show a study for the Beran estimator when the censoring indicators are not clearly specified, and next, propose a new method for the conditional survival function estimation with missing not a random (MNAR) censoring indicators. Along with the theoretical results, we illustrate how the estimators work for small samples by means of a simulation study and show their practical applicability with the analysis of synthetic data.

10:20 The mean, variance and correlation for bivariate recurrent event data with a terminal event
Frank Eriksson

Abstract: Recurrent events in the presence of a terminal event are often encountered in a biomedical setting. The marginal mean of the number of recurrent events in a specified time period is a useful non-parametric summary of recurrent events data also in the presence of a terminal event. Other useful non-parametric summaries, that are simple to compute, are the distribution function of the number of recurrent events for each point in time and the variance of the number of recurrent events. For bivariate recurrent events, still in the presence of a terminal event, we suggest a simple non-parametric estimator of the correlation of the marginal number of events for both processes. When there is no terminal event, the dependence in the processes can be described by this correlation, but with a non-negligable terminal event, interpretation is more involved. We suggest an adjustment for correlation induced by the terminal event to obtain a measure that reflects the dependence in the recurrent event processes among survivors only. Our estimators can be used for deciding whether the two recurrent events are correlated and in what way. We provide large sample properties of our estimators and show their performance in small samples by simulations. The estimators are applied in a study of catheter complications among patients receiving home parenteral nutrition through a central venous catheter, and we show a positive correlation between the number of infections and the number of occlusion defects.

10:40 Seeing the Unseen - a New Scheme for Missing Mass Estimation
Amichai Painsky

Abstract: Consider a finite sample from an unknown distribution over a countable alphabet. The missing mass refers to the probability of symbols that do not appear in the sample. Estimating the missing mass is a basic problem in statistics, machine learning and related fields, which dates back to the early work of Laplace, and the more recent seminal contribution of Good and Turing. The missing mass problem has many applications in a variety of domains. For example, given the observed population of Covid mutations, what is the probability that we encounter a new variant that has not been seen before? In this work we introduce a generalized framework for missing mass estimation. Our framework provides novel risk bounds and improves upon currently known estimation schemes. Importantly, it is easy to apply and does not require additional modeling assumptions. This makes it a favorable choice for many practical setups. Furthermore, we show that by utilizing our scheme, we improve the estimation accuracy of large alphabet probability distributions.

High-dimensional data

Chair: Nicolas Tavernier
Room: Akamas C

10:00 Exponential bounds for regularized Hotelling statistics in high dimension
El Mehdi Issouani, Patrice Bertail, Emmanuelle Gautherat

Abstract: In many applications (for instance in genomics or natural language processing), the dimension of the parameter of interest q is large in comparison to the sample size n and sometimes increasing with n. Consider for instance the problem of estimating or testing a mean of q-dimensional variables, with q>n; in that case, the empirical covariance matrix is not full rank and does not even converge to the true one when n goes to infinity, so that the usual "studentized statistics" or Hotelling-T2 tests are no longer valid. It is thus important to construct estimators and testing procedures which take into account the high dimensional aspects of the problem. One relevant proposition which has been developed in the statistical literature is to use a penalized estimator of the covariance matrix which is invertible and to use this matrix in tests. In that spirit Chen et al.(2011) have obtained asymptotically valid penalized Hotelling-T2 tests for the mean in a high dimension framework, when n and q=q(n) goes to infinity at some specific rate. The purpose of this talk is to further explore the finite sample properties of such tests by deriving exponential bounds of some correctly penalized Hotelling-T2. In this work, we obtain a Bernstein-type inequality for penalized self normalized sums (or penalized Hotelling-T2 that uses a penalized form of the covariance matrix) under very few moment conditions, for values that are large enough. This inequality depends on some parameters which in turn can be estimated yielding exponential inequalities with an additional explicit error term.

10:20 Quantile LASSO with changepoints in panel data models
Matus Maciak

Abstract: Panel data with change points are commonly used in all kinds of empirical problems under various regularity assumptions. We investigate the panel data models with structural breaks and the atomic pursuit techniques and nonparametric quantile estimation approaches are employed to construct the final estimate. Robust estimates and a complex insight into the data generating mechanism are both achieved by adopting the quantile LASSO approach. The final model is produced in a fully data-driven manner in just one single step. The final estimate is, under some reasonable assumptions, shown to be consistent with respect to the model estimation and the changepoint detection performance. The finite sample properties are investigated in a simulation study and the proposed methodology is applied for the Apple call option pricing problem.

10:40 Nonlinear shrinkage estimation of large-dimensional covariance matrices using splines
Nicolas Tavernier, Geert Dhaene

Abstact: The optimal, but infeasible, rotation-equivariant covariance matrix estimator under Frobenius loss shrinks each eigenvalue of the sample covariance matrix individually. We propose a feasible version thereof by approximating the highly nonlinear shrinkage function by penalized splines. The resulting estimator has a simple closed form and is, under very mild assumptions, asymptotically equivalent to its oracle counterpart when the sample size and the dimension grow at the same rate. Simulation experiments demonstrate the favorable performance of the proposed estimator compared to other rotation-equivariant estimators, including the linear shrinkage estimator of Ledoit and Wolf (JMVA 2004) and the state-of-the-art nonlinear shrinkage estimator of Ledoit and Wolf (AoS 2020). We also discuss applications in financial economics.

Copulas

Chair: Amitava Mukherjee

Room: Aphrodite A

10:00 Tests of constant conditional dependence structures over partition sets
Jean-David Fermanian, Alexis Derumigny, Aleksey Min

Abstract: Recently, the simplifying assumption for conditional copulas has been rigorously investigated and several testing procedures have been proposed. But dealing with numerous conditioning variables remains costly and often unfeasible under the usual perspective of pointwise conditioning events. To circumvent the curse of dimesion, we consider constant conditional dependence structures over some partition subsets. We propose a testing procedure of the constancy of conditional copulas w.r.t. such subsets, based on the equality of Kendall's rank correlation coefficients. This idea is extended to more general dependence measures by stating the weak convergence of conditional copula processes indexed by subsets of covariates. The performance of the proposed test will be illustrated in a simulation study and with real data.

10:20 Variable clustering using divergence type of dependence measures between random vectors
Steven De Keyser, Irène Gijbels

Abstract: Variable clustering aims at segmenting objects into homogeneous, separated groups and frequently serves as a crucial pre-processing step in model building and selection. Popular are hierarchical model-based architectures relying on similarities described by stochastic dependencies. Traditionally, these are derived from correlation or concordance measures and therefore restricted to the detection of bivariate monotonic relationships. We propose to use a class of margin-free (copula-based) divergence measures arising from concepts in information theory. When estimated non-parametrically, they detect any deviation from independence and hence succeed in finding distinct groups when other measures might fail. Moreover, the general theoretical framework considers the quantification of dependence between an arbitrary amount of random vectors, meaning that aggregation criteria are not needed as such when the groups are composed of multiple variables. We investigate properties of the considered dependence measures and discuss their estimation, with specific attention to non-parametric density estimation on the copula support. Finally, we illustrate their practical use in simulation studies and real data examples, with particular focus on rank-invariant clustering of continuous data. This talk is based on joint work with I. Gijbels.

10:40 Nonparametric Surveillance of Unknown Structural Dependence in High-Dimensional Processes using Pseudo Copula and Eigenvectors – A Computational Approach
Amitava Mukherjee

Abstract: This presentation considers nonparametric online monitoring of unknown structural dependence in multivariate and high-dimensional data streams using pseudo copula observation and the eigenvector of a suitably estimated covariance matrix. The paper presents a simplified computational approach. We study the in-control robustness, effect of estimation of covariance matrix in the high-dimensional setup and investigate the proposed scheme's out-of-control performance in various situations. Some comparisons with the existing procedures based on Monte-Carlo indicate that the proposed design is more robust, efficient and flexible. We consider an application in quality monitoring semiconductor manufacturing data.

Screening and classification

Chair: Dragan Radulovic

Room: Aphrodite B

10:00 Novel nonparametric specification tests for additive concurrent model formulation
Laura Freijeiro-González, Manuel Febrero-Bande, Wenceslao González-Manteiga

Abstract: Novel nonparametric specification tests are developed for the additive concurrent model formulation to assess covariates selection in the concurrent model. As a result, they provide tools to determine if there exists any general additive effects between covariates and the response, allowing to implement covariates screening in the concurrent model framework. In contrast with other proposals in literature, these have the advantage that there is not need of model or smoothing parameters preliminary estimation, resulting in nonparametric alternatives. For this purpose, we make use of new adaptations of the well-known covariance distance. An example is the martingale difference divergence coefficient of Shao and Zhang (2014). We introduce new global dependence tests to quantify the effect of a subset of covariates in the response, jointly with partial dependence tests to apply covariates selection. In both cases, the whole observed time instants information is taken under consideration in the statistics construction. Their asymptotic distribution is obtained on each case and a bootstrap algorithm is proposed to obtain the p-value in practice. Furthermore, their good performance is illustrated by means of simulations as well as their application to real datasets. Besides, their behaviour is compared with popular competitors in literature for both: linear and non-linear concurrent models structure.

10:20 Nonparametric Screening and Selection in presence of dependence among predictors with applications to environmental pollutants
Sanjib Basu

Abstract: Variable selection under dependence among predictors is well studied but remains challenging. We develop a model free screening and selection method. This method is based on the distance correlation, a nonparametric measure of dependence in arbitrary dimensions, whose null value imply independence, and whose sure independence screening property has been studied. Our proposed method considers both screening and selection and additionally addresses selection in presence of strong dependence among predictors. We compare performances with existing methods under linear and nonlinear data generating models with moderate to strong collinearity. Our motivating application arises in environmental epidemiology where individuals are exposed simultaneously to a multitude of pollutants in the environmental mixture which potentially interact and presents health risk. The pollutant measures are often strongly correlated at levels that are generally not seen in other areas of science. The proposed method is applied in assessing effect of environmental pollutants on diabetes health outcomes from the US National Health and Nutrition Examination Survey.

10:40 Accelerated Random Search for Weighted Nearest Neighbors
Dragan Radulovic

Abstract: Classification problem is as follows: Given two sets of d-dim data, {X} and {Y} and a new point Z, determine if Z belongs to X or to Y cluster. Classical methods are heavily parametric and most of them rely on Gaussian assumptions. Among non-parametric techniques, The Nearest Neighbor method is arguably the most commonly used. Its modification, The Weighted Nearest Neighbors, puts different weights at different coordinates, and consequently it is much more powerful method. Unfortunately, computing the right weights is exceedingly difficult optimization problem. Here we apply Accelerated Random Search technique to optimize the weights. We present the overview of the method and offer numerous computational results where we contrast its performance with standard methods.

Ranks and permutations

Chair: Niels Olsen

Room: Leda

10:00 A comparison of statistical methods for analyzing longitudinally measured ordinal outcomes in rare diseases
Martin Geroldinger, Johan Verbeeck, Konstantin Emil Thiel, Geert Molenberghs, Arne Bathke, Georg Zimmermann

Abstract: Ordinal data in a repeated measures design of a cross-over study for rare diseases usually do not allow for the use of standard parametric methods. Hence, non-parametric methods should be considered instead. However, due to the complexity of the design, choosing an appropriate nonparametric approach is difficult, too, and only limited simulation studies for settings with very small sample sizes exist. Therefore, starting from an epidermolysis bullosa simplex trial with the above-mentioned design, a rank-based approach using the R package nparLD and different generalized pairwise comparisons (GPC) methods were compared neutrally in a simulation study. The results revealed that there was not one single best method for this particular design, since a trade-off between achieving high power, accounting for period effects, and missing data must be made. Specifically, nparLD does not address cross-over aspects, and the GPC variants partly ignore the longitudinal information. Overall, the prioritized unmatched GPC method achieved the highest power in the simulations scenarios, although this may be due to the specified prioritization. The rank-based approach yielded good power even at a sample size of N=6, while the matched GPC method could not control the type I error. Together with the results from extensive simulation studies using binary and count outcome data, the findings of the present contribution will constitute the basis for developing recommendations and educational materials that will be disseminated in the statistical as well as in the clinical scientific community and should contribute to accurateness of methodological approaches of clinical research in rare diseases.

10:20 Anomaly Detection for a Large Number of Streams: A Permutation/Rank-Based Higher Criticism Approach
Ivo V. Stoepker, Rui M. Castro, Ery Arias-Castro, Edwin R. van den Heuvel

Abstract: Anomaly detection when observing a large number of data streams is essential in a variety of applications, ranging from epidemiological studies to monitoring of complex systems. Detection is commonly cast as a hypothesis testing problem. High-dimensional scenarios are usually tackled with scan-statistics and related methods, requiring stringent modeling assumptions for proper test calibration. In this talk we take a non-parametric stance. We propose two variants of the higher criticism statistic not requiring knowledge of the null distribution, either through the use of permutations, or through data ranks. Both methodologies result in an exact test in finite samples. Our permutation methodology is applicable when observations within null streams are independent and identically distributed, and we show this methodology is asymptotically optimal in the wide class of exponential models. Our rank-based methodology is more flexible, and only requires observations within null streams to be independent. We show this use of ranks incurs an asymptotic power loss which we can fully characterize based on the probability of mis-ranking a null observation. For many common models, this asymptotic power loss is minimal. For both methodologies, we demonstrate the power loss in finite samples is minimal with respect to the oracle test. Furthermore, since the proposed statistics do not rely on asymptotic approximations, they typically perform better than popular variants of higher criticism that rely on such approximations. We include recommendations such that the tests can be readily applied in practice, including adjustments to make them appropriate when there is some degree of correlation within streams, and demonstrate its applicability in monitoring the daily number of COVID-19 cases in the Netherlands.

10:40 Confidence regions for univariate and multivariate data using permutation tests
Niels Olsen

Abstract: Confidence intervals are central to statistical inference as a tool to evaluate the type I error risk at a given significance level. We present a novel method to construct non-parametric confidence intervals using only a single run of a permutation test. This methodology is extended to a multivariate setting, where we are able to handle multiple testing under arbitrary dependence. We demonstrate the method on a weather data set and in a simulation example.

Functional & Clustered Data

Chair: Tomas Masak

Room: Athena

10:00 Convolutional Networks and Tensor Product ANOVA Smoothing Splines
Zaid Harchaoui, Meyer Scetbon

Abstract: We present a description of the function space and the smoothness class associated with a convolutional network using the machinery of reproducing kernel Hilbert spaces. We show that the mapping associated with a convolutional network expands into a sum involving elementary functions akin to spherical harmonics. This functional decomposition can be related to the functional ANOVA decomposition in nonparametric statistics. Building off our functional characterization of convolutional networks, we obtain statistical bounds highlighting an interesting trade-off between the approximation error and the estimation error.

10:20 Nonparametric regression with clustered observations
Stanislav Anatolyev

Abstract: We consider a nonparametric mean regression for clustered samples, where observations are independent across clusters, but may exhibit within-cluster dependence and be accompanied with conditional heteroskedasticity. The clusters may have different size, and their average size may be small, moderately large, or seriously large. We focus on the Nadaraya-Watson (NW) mean regression estimator, and derive its asymptotic distribution under three corresponding scenarios of a fixed, slowly growing, and rapidly growing average cluster size, which interplays with the rate of bandwidth shrinkage, as the total sample size increases. While in all three cases the NW estimator is asymptotically normal and the asymptotic bias has similar forms, the rate of convergence and asymptotic variance sharply differ depending on whether the average cluster size is fixed or slowly growing, or whether it is growing fast. We also discuss the issue of optimal bandwidth selection based on the standard criterion of asymptotic mean squared error; this rule turns out to be different in the two cases as well. The form of asymptotic variance of the NW estimator depends on the rate of the average cluster size via the dominance or balance between within-cluster error variances and error covariances. The former dominate when the average cluster size grows slowly enough, while the latter dominate when the growth is fast. In the intermediate case, when both error variances and covariances are balanced, the asymptotic variance of the NW estimator has two components. We propose a number of asymptotic variance estimates, which are suitable depending on the rate of growth of the average cluster size, and prove their consistency. One of these estimators turns out to be robust to this growth rate, and so can be used universally without resorting to the question of how big the rate is qualified in an actual application

10:40 Inference and Computation for Sparsely Sampled Random Surfaces
Tomas Masak

Abstract: Non-parametric inference for functional data over two-dimensional domains entails additional computational and statistical challenges, compared to the one-dimensional case. Separability of the covariance is commonly assumed to address these issues in the densely observed regime. Instead, we consider the sparse regime, where the latent surfaces are observed only at few irregular locations with additive measurement error, and propose an estimator of covariance based on local linear smoothers. Consequently, the assumption of separability reduces the intrinsically four-dimensional smoothing problem into several two-dimensional smoothers and allows the

proposed estimator to retain the classical minimax-optimal convergence rate for two-dimensional smoothers. Even when separability fails to hold, imposing it can be still advantageous as a form of regularization. A simulation study reveals a favorable bias-variance trade-off and massive speed-ups achieved by our approach. Finally, the proposed methodology is used for qualitative analysis of implied volatility surfaces corresponding to call options, and for prediction of the latent surfaces based on information from the entire data set, allowing for uncertainty quantification. Our cross-validated out-of-sample quantitative results show that the proposed methodology outperforms the common approach of pre-smoothing every implied volatility surface separately.

Time series

Chair: Carina Beering
Room: Zeus

10:00 Fully data-driven non-parametric estimation of Toeplitz covariance matrices
Karolina Klockmann

Abstract: Estimation of the Toeplitz covariance matrix of a vector with a known mean is a central problem in many areas of multivariate analysis and is a complex task. In this context two main frameworks are considered: either there are N independent and identically distributed p-dimensional random vectors with a Toeplitz covariance matrix, or there is a single realization of a stationary stochastic process, given as one N-dimensional vector. In both frameworks the matrix of the sample covariances is known to be inconsistent in the spectral norm, so that regularized versions, such as the tapering or banding covariance estimators, have been proposed. The minimax optimality of the tapering estimator has been shown for the Toeplitz covariance matrices with a certain decay of their covariance sequence and for those with a sufficiently smooth spectral density function. However, such estimators are not guaranteed to be positive definite and data-driven choices of the regularization parameters are not available. In this talk we present an estimator for the Toeplitz covariance matrix and its inverse which overcome these drawbacks. First, we derive an alternative version of the Whittle likelihood based on the Discrete Cosine Transform matrix, which is shown to asymptotically diagonalize Toeplitz matrices. This leads to the problem of spectral density estimation in a Gamma regression setting. Using variance stabilizing transforms of Brown, Cai and Zhou (2010) the transformed spectral density can be estimated with a periodic smoothing spline in a Gaussian regression setting. The resulting estimators for the Toeplitz covariance matrix and its inverse are positive definite and all regularization parameters are data-driven. As our main result, we show that our estimators attain the minimax optimal convergence rate. The performance of our estimators is compared to regularized versions of the sample covariance matrix in a simulation study.

10:20 A Test of Independence for Locally Stationary Processes Using a Weighted Characteristic Function- based Distance
Carina Beering

Abstract: We propose a testing procedure for independence of locally stationary processes using a weighted distance composed of characteristic functions (CF) and its empirical version as a base. The distance covariance defined by Székely et al. (2007) and its use by Jentsch et al. (2020) inspired the essential idea of this concept. To be finally able to compile a testing procedure, we provide the needed results with the notion of the beneficial effects of a bootstrap analogue. Therefore, bootstrap versions of the previously presented findings are established. Beforehand, the concept of empirical weighted CF-distance is transferred to the bootstrap world. In the end, we perform a simulation study which uses our testing procedure to detect independence as well as dependence of different forms.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Special Invited Talk: Victor Chernozhukov

Chair: Ingrid Van Keilegom

Room: Akamas A

Long Story Short: Omitted Variable Bias in Causal Machine Learning

Abstract: We derive general, yet simple, sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many of the traditional targets of investigation in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, and policy effects from shifts in covariate distribution -- all for general, nonparametric causal models. Our construction relies on the Riesz-Frechet representation of the target functional. Specifically, we show how the bound on the bias depends only on the additional variation that the latent variables create both in the outcome and in the Riesz representer for the parameter of interest. Moreover, in many important cases (e.g, average treatment effects and average derivatives) the bound is shown to depend on easily interpretable quantities that measure the explanatory power of the omitted variables. Therefore, simple plausibility judgments on the maximum explanatory power of omitted variables (in explaining treatment and outcome variation) are sufficient to place overall bounds on the size of the bias. Furthermore, we use debiased machine learning to provide flexible and efficient statistical inference on learnable components of the bounds. Finally, empirical examples demonstrate the usefulness of the approach.

12:30 - 13:30 Lunch Break

16:00 - 22:00 Excursion & Conference Dinner

THURSDAY 23 JUNE 2022

9:00 - 11:00 Invited Paper Session 6

New issues in deconvolution problems

Organiser: Clément Marteau

Chair: Clément Marteau

Room: Akamas A

9:00 Adaptive minimax testing for circular convolution
Sandra Schluttenhofer, Jan Johannes

Abstract: Given observations from a circular random variable contaminated by an additive measurement error, we consider the problem of minimax optimal goodness-of-fit testing in a non-asymptotic framework. We propose direct and indirect testing procedures using a projection approach. The structure of the optimal tests depends on regularity and ill-posedness parameters of the model, which are unknown in practice. Therefore, adaptive testing strategies that perform optimally over a wide range of regularity and ill-posedness classes simultaneously are investigated. Considering a multiple testing procedure, we obtain adaptive i.e. assumption-free procedures and analyse their performance. Compared with the non-adaptive tests, their radii of testing face a deterioration by a log-factor. We show that for testing of uniformity this loss is unavoidable by providing a lower bound. The results are illustrated considering Sobolev spaces and ordinary or super smooth error densities.

9:30 Deconvolution for some singular density errors with respect to the L1 loss
Mathieu Sart

Abstract: We are interested in the problem of estimating a density from noisy observations. The deconvolution problem is assumed to be moderately ill-posed with small degree of ill-posedness. We propose a versatile, model-based, estimation procedure. We use it to get non-asymptotic risk bounds for the L1 loss under regularity, shape, or parametric constraints on the density. More precisely, we study the situation where the density belongs to a possibly inhomogeneous Besov space. This leads to new upper-bounds of the minimax risks on Besov semi-balls. We then consider the situation where the density is compactly supported and monotone on its support. More generally, the density may be multimodal (with possibly unknown modes). We also explain how to replace the monotone assumption by a convexity one. Moreover, the rates we get become almost parametric whenever the density is piecewise constant (or affine) on a known number of unknown intervals. Finally, we cope with the situation where the density is known, up to location and scale parameters. In such a model, the traditional maximum likelihood method does not work, unlike ours. Besides, our estimator is robust with respect to model misspecification.

10:00 Rate optimal estimation of quadratic functionals in inverse problems with partially known operator
Martin Kroll

Abstract: We consider rate optimal estimation of a quadratic functional in an inverse Gaussian sequence space model. We assume that the operator of the inverse problem is already given in diagonal form but the sequence of eigenvalues is not known but only accessible through additional observations. We derive optimal rates of convergence for this problem with special emphasis on the impact of the perturbation in the eigenvalues. We also discuss implications of the derived results for the theory of signal detection and goodness-of-fit testing in inverse problems. Some potential directions for future research will be indicated to conclude the talk.

Statistics under nonstationarity

Organiser: Anne Leucht

Chair: Anne Leucht

Room: Akamas C

9:00 Poisson Network Autoregression
Konstantinos Fokianos, Mirko Armillotta

Abstract: We consider network autoregressive models for count data with a non-random neighborhood structure. The main methodological contribution is the development of conditions that guarantee stability and valid statistical inference for such models. We consider both cases of fixed and increasing network dimension and we show that quasi-likelihood inference provides consistent and asymptotically normally distributed estimators. The work is complemented by simulation results and a data example.

9:30 Bootstrap for integer-valued GARCH(p,q) processes
Michael Neumann

Abstract: We consider integer-valued processes with a linear or nonlinear GARCH structure, where the count variables given the past follow a Poisson distribution. We show that a contraction condition imposed on the intensity function yields a contraction property of the Markov kernel of the process. This allows almost effortless proofs of the existence and uniqueness of a stationary distribution as well as of absolute regularity of the count process. As our main result, we construct a coupling of the original process and a model-based bootstrap counterpart. Using a contraction property of the Markov kernel of the coupled process we obtain bootstrap consistency for different types of statistics.

10:00 Bootstrap confidence bands for the mean and covariance kernel of Banach space valued functional data
Melanie Birke, Christoph Reihl

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

10:30 Learning to reflect
Lukas Trottner

Abstract: Theoretical solutions to stochastic optimal control problems are well understood in many scenarios, however their practicability suffers from the assumption of known dynamics of the underlying stochastic process. This raises the challenge of developing purely data-driven strategies. In this talk we focus on a particular ergodic ergodic control problem for Lévy processes, whose theoretical solution calls for statistical estimation of an optimal reflection boundary given as the maximizer of a generator functional of the associated ascending ladder height subordinator. Even though this process has desirable statistical properties, a direct plug-in approach based on nonparametric estimation of its Lévy triplet is not feasible since even with a full record of the parent Lévy process at hand we cannot recover the paths of its ascending ladder height process. We solve this problem by considering an appropriate space/time transformation of the Lévy process to obtain a directly observable spatially ergodic process that reveals all necessary statistical information. This spatial structure allows us to construct a temporal estimator of the optimal reflection boundary and to obtain non-asymptotic rates for its regret based on general concentration results for Lévy processes and additive functionals of Markov processes.

Topics in Econometrics: partial identification, boundary-adaptive kernel density estimation, and more

TBA

Organiser: Jeff Racine

Chair: Jeff Racine

Room: Aphrodite A

09:00 Estimating High Dimensional Monotone Index Models by Iterative Convex Optimization1
Shakeeb Khan, Xiaoying Lan, Elie Tamer

Abstract: In this paper we propose a new approach to estimating large dimensional monotone index models. This class of models has been popular in the applied and theoretical econometrics literatures as they include discrete choice, nonparametric transformation, and duration models. The main advantage of our approach is computational: in comparison, rank estimation procedures such as proposed in Han (1987) and Cavanagh and Sherman (1998) optimize a nonsmooth, non convex objective function, and finding a global maximum gets increasingly di cult with a large number of regressors. This makes such procedures particularly unsuitable for “big data” models. For our semiparametric model of increasing dimension, we propose a new algorithm based estimator involving the method of sieves and establish asymptotic its properties. The algorithm uses an iterative procedure where the key step exploits its strictly convex objective function. Our main results here generalize those in, e.g. Dominitz and Sherman (2005) and Toulis and Airoldi (2017), who consider algorithmic based estimators for models of fixed dimension.

On Optimal Set Estimation for Partially Identified Binary Choice Models
Shakeeb Khan, Tatiana Komorova, Denis Nekipelou

Abstract: In this paper we reconsider the notion of optimality in estimation of partially identified models. We illustrate the problem in the context of a semiparametric binary choice model with discrete covariates as an example of a model which is partially identified as shown in, e.g. Bierens and Hartog (1988). A set estimator for the regression coe cients in the model can be constructed by implementing the Maximum Score procedure proposed by Manski (1975). For many designs this procedure converges to the identified set for these parameters, and so in one sense is optimal. But as shown in Komarova (2013) for other cases the Maximum Score procedure converges to an outer region of the identified set. This motivates alternative methods that are optimal in the sense that they converge to the identified region in all designs, and we propose two such procedures. One is a Hodges type estimator combining the Maximum Score estimator with existing two step procedures. A second is a two step estimator using a Maximum Score type objective function in the second step. We show the Hodges type estimator has poor “uniformity” properties by allowing for drifting parameter sequences. Furthermore, by taking this notion of uniformity into account, the optimality of both the Maximum Score and two-step estimators can be established. This last result is loosely analogous to the optimality of MLE among the class of regular estimators.

09:30 Dynamic Ordered Panel Logit Models
Chris Muris, Bo Honoré, Martin Weidner

Abstract: We study a dynamic ordered logit model for panel data with fixed effects. We establish the validity of a set of moment conditions that are free of the fixed effects and that can be computed using four or more periods of data. We establish sufficient conditions for these moment conditions to identify the regression coefficients, the autoregressive parameters, and the threshold parameters. The parameters can be estimated using the generalized method of moments. We document the performance of this estimator using Monte Carlo simulations and an empirical illustration to self-reported health status using the British Household Panel Survey.

10:00 Time-Varying Linear Transformation Models with Fixed Effects and Endogeneity for Short Panels
Irene Botosaru, Chris Muris, Senay Sokullu

Abstract: This paper considers a class of fixed-T nonlinear panel models with time-varying link function, fixed effects, and endogenous regressors. We establish sufficient conditions for the identification of the regression coefficients, the time-varying link function, the distribution of the counterfactual outcomes, and certain (time-varying) average partial effects. We propose estimators for these objects and study their asymptotic properties. We show the relevance of our model by estimating the effect of teaching practices on student attainment as measured by test scores on standardized tests in mathematics and science. We use data from the Trends in International Mathematics and Science Study, and show that both traditional and modern teaching practices have positive effects of similar magnitudes.

10:30 Boundary-Adaptive Kernel Density Estimation - Empirical Support Kernels and Cross-Validation: The Curious Case of the Uniform Density
Jeffrey Racine, Qi Li

Abstract: We investigate an extra-optimality property present when we model bounded-support densities using appropriate kernel methods with data-driven bandwidth selection and empirical (as opposed to known) support bounds. We can observe, theoretically and in finite-sample settings, that the kernel method is capable of outperforming even correctly specified parametric models, which we term here the extra-optimality of the proposed approach. In particular, the closer the underlying density is to a uniform distribution on [min(x),max(x)] the closer is the proposed method to the unknown true density not its correctly specified parametric estimate peer. We demonstrate that this result has implications when modelling a range of densities and not only the uniform. This suggests that appropriate methods for kernel density estimation in bounded-support settings can achieve much faster rates of convergence than generally known, and can even dominate those of correctly specified parametric models hence can improve upon even correctly specified parametric models in finite-sample settings.

New developments in rank-based methods

Organiser: Michele La Rocca
Chair: Valeria Vitelli
Room: Aphrodite B

09:00 Avoiding Degeneracies in Ordinal Unfolding Using Kemeny-Equivalent Dissimilarities for Two-Way Two-Mode Preference Rank Data
Antonio D'Ambrosio

Abstract: A simple but effective procedure to avoid degeneracies in ordinal Unfolding for preference rank data based on the Kemeny distance is proposed. Considering Unfolding as a particular MDS procedure with missing within-set proximities, unknown proximities are first estimated using correlations related to the Kemeny distance, and then the complete proximity matrix is analyzed in a standard MDS framework. A simulation study shows that our proposal is able to both recover the order of the preferences and reproduce the position of both rankings and objects in a reduced geometrical space

09:30 Permutation testing for thick data when the number of variables is much greater than the sample size
Patrick Langthaler, Riccardo Ceccato, Arne Bathke, Rosa Arboretti, *Luigi Salmaso*

Abstract: Non-parametric combination (NPC) of p-values provides a useful approach for testing multivariate equality of distributions. Several different combining functions have been suggested for example by Fisher, Stouffer and Tippett. There is however a lack of recommendations which combining function is best under which circmustances for nonparametric testing. We present an intensive simulation study, assessing the performance of different combining functions and combination procedures for one-factorial designs with the number of variables much larger than the number of observations. Examined settings include two-sample problems, a C-sample problem and a stochastic ordering problem with variables following different continuous distributions or a mix of continuous and discrete distributions with different correlation structures. In all settings there is a fixed number of informative variables and an increasing number of noninformative variables. We explore the performance of the NPC with different combining functions compared to rank- and distance-based competing procedures.

10:00 Ranks or Pseudo-Ranks? Some practical considerations
Georg Zimmermann, Arne C. Bathke

Abstract: Nonparametric rank-based methods are frequently used in applied research: On the one hand, nonparametric approaches are often applied when the assumptions underlying a particular parametric test are violated. On the other hand, rank-based methods are also

applicable to ordinal data. Therefore, they may be used for analyzing quality of life scores, disease severity scales, or more generally, any rating or score. Examples of commonly used rank-based tests are the Wilcoxon-Mann-Whitney test and the Kruskal-Wallis test. However, it has been shown recently that some rank-based methods might yield paradoxical results in particular settings. As a solution to this problem, so-called "pseudo-ranks" have been introduced, which have the advantage of not being affected by those paradoxa. Nevertheless, the pseudo-rank-based approach leads to a different definition of the relative effect and, thus, also to a different interpretation. Therefore, the aim of this talk is to give the reader some guidance regarding the interpretation of the results obtained from nonparametric (pseudo-) rank-based tests, with special emphasis being placed on potential pitfalls as well as on some consequences of recent theoretical work for appropriately using these methods in applied research.

10:30 Pseudo-Mallows for Efficient Probabilistic Preference Learning
Valeria Vitelli

Abstract: We propose the Pseudo-Mallows distribution over the set of all permutations of n items, to approximate the posterior distribution with a Mallows likelihood. The Mallows model has been proven to be useful for recommender systems where it can be used to learn personal preferences from highly incomplete data provided by the users. Inference based on MCMC is however slow, preventing its use in real time applications. The Pseudo-Mallows distribution is a product of univariate discrete Mallows-like distributions, constrained to remain in the space of permutations. The quality of the approximation depends on the order of the n items used to determine the factorization sequence. In a variational setting, we optimise the variational order parameter by minimising a marginalized KL-divergence. We propose an approximate algorithm for this discrete optimization, and conjecture a certain form of the optimal variational order that depends on the data. Empirical evidence and some theory support our conjecture. Sampling from the Pseudo-Mallows distribution allows fast preference learning, compared to alternative MCMC based options, when the data exists in form of partial rankings of the items or of clicking on some items. Through simulations and a real life data case study, we demonstrate that the Pseudo-Mallows model learns personal preferences very well and makes recommendations much more efficiently, while maintaining similar accuracy compared to the exact Bayesian Mallows model.

New frontiers in network data analysis

Organiser: Marianna Pensky & Srijan Sengupta
Chair: Srijan Sengupta
Room: Christian Barnard

09:00 Clustering of Diverse Multiplex Networks
Marianna Pensky

Abstract: The talk introduces the DIverse MultiPLEx (DIMPLE) network model where all layers of the network have the same collection of nodes and are equipped with the Stochastic Block Models (SBM). In addition, all layers can be partitioned into groups with the same community structures, although the layers in the same group may have different matrices of block connection probabilities. The DIMPLE model generalizes a multitude of papers that study multilayer networks with the same community structures in all layers (which include the tensor block model and the checker-board model as particular cases), as well as the Mixture
Multilayer Stochastic Block Model (MMLSBM), where the layers in the same group have identical matrices of block connection probabilities. Since the techniques from either of the above mentioned groups cannot be applied to the DIMPLE model, we introduce novel algorithms for the between-layer and the within-layer clustering.

9:30 Avoiding a popularity contest: network clustering with the random walk Laplacian
Alexander Modell

Abstract: Community detection on networks often follows a two-step procedure: embedding, in which nodes are represented as points in space, and subsequent clustering. Often, it is of interest to cluster nodes in a way which is agnostic to degree. In this talk, I will discuss spectral clustering with the random walk Laplacian matrix, and show that it satisfies this criterion by representing nodes on a projective plane. Theoretical results, such as uniform consistency and a central limit theorem motivate clustering the embedding using a weighted Gaussian mixture model, which allows simultaneous estimation of the embedding dimension and the number of clusters using the Bayesian Information Criterion (BIC). Ideas are illustrated using light-hearted examples, such as a network of enmities between characters of the Harry Potter books.

10:00 Pseudo-likelihood-based M-estimation of networks with dependent connections and parameter vectors of increasing dimension
Michael Schweinberger, Jonathan R. Stewart

Abstract: An important question in statistical network analysis is how to estimate models of dependent network data with intractable likelihood functions, without sacrificing computational scalability and statistical guarantees. We demonstrate that scalable estimation of random graph models with dependent edges is possible, by establishing consistency results and convergence rates for pseudo-likelihood-based M- estimators for parameter vectors of increasing dimension, based on a single observation of dependent random variables. The main results cover discrete Markov random fields with parameter vectors of increasing dimension in single-observation scenarios. To the best of our knowledge, these are the first such results. To showcase consistency results and convergence rates, we introduce a novel class of generalized β-models with dependent edges and parameter vectors of increasing dimension. We establish consistency results and convergence rates for pseudo-likelihood-based M-estimators of generalized β-models with dependent edges, in dense- and sparse-graph settings.

Advances in semiparametric and nonparametric copula-based inference

Organiser: Irène Gijbels
Chair: Irène Gijbels
Room: Leda

9:00 **Bernstein estimators for conditional copulas and their derivatives**
Noel Veraverbeke

Abstract: The dependence structure between two variables Y1 and Y2 can be highly influenced by a third variable X, called covariate. A complete description of the dependence of Y1 and Y2 given X = x is given by the so called conditional copula function Cx . The existing nonparametric estimators for Cx involve smoothing in the covariate variable x but are not smooth in the other variables. In this talk we study new versions of these estimators where the smoothing is done by the Bernstein method. This method is known to produce estimators for functions on the unit square with good bias and variance properties. We establish these asymptotic properties as well as asymptotic normality. We also deal with estimation of the partial derivatives of Cx in view of an application to risk ratios with covariates.

9:30 **A universal representation of statistical dependence**
Gery Geenens

Abstract: The concept of “dependence” is central to statistics. It is, therefore, noteworthy that it is rarely defined formally in the scientific literature. A view commonly held is that dependence can be simply defined as just “non-independence”. It is then not clear what the numerous dependence measures proposed in the literature really attempt to quantify, as measuring dependence seems incompatible with its alleged binary nature. It so appears that the way statistical dependence is actually handled is inconsistent with its assented definition. This misalignment has permitted a certain level of looseness when addressing dependence-related questions, causing numerous incongruities, the most obvious being that the most famous “measure of dependence”, namely Pearson’s correlation, may be null in the case of two variables deterministically bound hence does not characterise dependence at all. Arguing that a topic of this importance would benefit from slightly more rigid guidelines, in this work we attempt to provide a deeper understanding of the concept of dependence. Starting from a very simple definition of what should be understood by “dependence”, a cascade of implications follows and ultimately allows the identification of a universal representation of the dependence structure of a random vector. Some popular measures of dependence align with that representation, many others do not. We will also discuss the role of copulas from that perspective, showing that copulas provide a sensible approach for analysing and modelling dependence in a continuous vector but cannot be justified in a discrete setting.

10:00 **Kernel estimation of copulas for circular distributions**
Jose Ameijeiras-Alonso, Irène Gijbels

Abstract: On several occasions, we are interested in realizations of random angles, whose support is the unit circle. For example, when wind directions are measured at two different locations or in the study of the directions of movements of two different animal herds. In those situations, it may be also of interest to study the relation between these two random angles. When data lies on the torus, the analog of copulas for circular data is called circulas. In this talk, we will provide a new way to estimate these circulas using circular kernels. We will see how to obtain the asymptotic mean integrated squared error, from which the optimal smoothing parameter is derived. Also, different strategies for obtaining a plug-in smoothing parameter will be discussed. The applicability of these nonparametric circulas will be shown on a real dataset.

10:30 **Flexible multivariate distributions with central symmetry**
Jonas Baillien, Irène Gijbels, Anneleen Verhasselt

Abstract: Knowing if symmetry is present in data can help you making more accurate assumptions and in term improve results. Be it e.g. through more accurate estimation of the center or choosing more suited models. Univariate symmetry is a widely known and deeply explored topic, but in a multivariate setting, the notion of symmetry is less clear as different views exist. We will focus on multivariate central symmetry. We test for this using a generalized likelihood ratio test which can be used for parametric, semi-parametric and non-parametric models. In this we use families of distributions which are asymmetric by nature, but also contain their symmetric counterpart. A key focus is on skew-elliptical copulas, in particular the skew-normal copula. The concepts provided are demonstrated on real life data and the performance of tests for symmetry are empirically checked through simulation studies.

Nonparametric statistics and geometry

Organiser: Athanasios Georgiadis
Chair: Athanasios Georgiadis
Room: Athena

9:00 **Nonparametric estimation of covariance and autocovariance operators on the sphere**
Alessia Caponera, Julien Fageot, Matthieu Simeoni, Victor M. Panaretos

Abstract: We propose nonparametric estimators for the second-order central moments of spherical random fields within a functional data context. We consider a measurement framework where each field among an identically distributed collection of spherical random fields is sampled at a few random directions, possibly subject to measurement error. The collection of fields could be i.i.d. or serially dependent. Though similar setups have already been explored for random functions defined on the unit interval, the nonparametric estimators proposed in the literature often rely on local polynomials, which do not readily extend to the (product) spherical setting. We therefore formulate our estimation procedure as a variational problem involving a generalized Tikhonov regularization term. The latter favours smooth covariance/autocovariance functions, where the smoothness is specified by means of suitable Sobolev-like pseudo-differential operators. Using the machinery of reproducing kernel Hilbert spaces, we establish representer theorems that fully characterise the form of our estimators. We determine their uniform rates of convergence as the number of fields diverges, both for the dense(increasing number of spatial samples) and sparse (bounded number of spatial samples) regimes. We moreover validate and demonstrate the practical feasibility of our estimation procedure in a simulation setting, assuming a fixed number of samples per field. Our numerical estimation procedure leverages the sparsity and second-order Kronecker structure of our setup to reduce the computational and memory requirements by approximately three orders of magnitude compared to a naive implementation would require.

9:30 **Density estimation on manifolds or more general metric spaces**
Galatia Cleanthous, A..G. Georgiadis, G. Kerkyacharian, P. Petrushev, D. Picard

Abstract: We study the problem of density estimation on a general class of metric spaces, that unifies Euclidean space, sphere, ball, manifolds and more. Precisely, we derive oracle inequalities and upper bounds for kernel and wavelet density estimators. Moreover we extract lower bounds, that are highly effected by the space geometry.

10:00 **On high-frequency limits of U-statistics in Besov spaces over compact manifolds**
Claudio Durastanti, Domenico Marinucci, Xiaohong Lan

Abstract: In this talk, quantitative bounds in high-frequency central limit theorems are derived for Poisson based U-statistics of arbitrary degree built by means of wavelet coefficients over compact Riemannian manifolds. The wavelets considered here are the so-called needlets, characterized by strong concentration properties and by an exact reconstruction formula. Furthermore, we consider Poisson point processes over the manifold such that the density function associated to its control measure lives in a Besov space. We will discuss new rates of convergence, depending strongly on the degree of regularity of the control measure of the underlying Poisson point process, providing a refined understanding of the connection between regularity and speed of convergence in this setting. This is a joint work with Solesne Bourguin

10:30 **On the rate of estimation for the invariant density of stochastic differential equations.**
Chiara Amorino, Arnaud Gloter

Abstract: In this talk, we will discuss some results on the estimation of the invariant density associated to a multivariate diffusion X=(X-_t)_{t \ge 0}, solution of stochastic differential equations. The estimation of the invariant density is a problem of great relevance because of the huge amount of applications in physics and numerical methods, the Markov Chain Monte Carlo above all. Evidence of the attractiveness of the non-parametric estimation for the stationary measure of a continuous mixing process is the fact such a subject is both a long-standing problem and a living topic. We propose kernel density estimators and we measure their accuracy by studying the size of their pointwise L^2 error. We first of all find the convergence rates associated to the proposed estimators. After that, we wonder if it possible to propose other estimators which achieve better convergence rates.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Keynote Talk: Martin Wainwright
Chair: Alexander Goldenshluger
Room: Akamas A

Non-parametric estimation for reinforcement learning

Abstract: Reinforcement learning (RL) consists of statistical procedures that use data to determine near-optimal policies for sequential decision-making. It makes use of Markov decision processes, and has applications in numerous areas (e.g., environmental control, robotics, supply chain management, competitive game-playing, industrial process control). Non-parametric methods have a central role to play in this setting, but the RL setting brings several challenges that are not present for “static” non-parametric estimation problems, including the role of the Markovian dynamics and the challenges of off-policy data. In this talk, we provide an overview of these challenges, and discuss some recent progress, including non-asymptotic guarantees for procedures based on kernel methods, as well as non-parametric forms of Bellman residual and projection methods. Based on joint research with: Yaqi Duan, Mengdi Wang and Andrea Zanette.

12:30 - 13:30 Lunch Break

13:30 - 15:30 Invited Paper Session 7

Nonparametric Testing

Organiser: Miguel Delgado

Chair: Miguel Delgado

Room: Akamas A

13:30 On the T-test
Sergei. Novak

Abstract: We show that the T-test can be misleading. We argue that normal or Student's approximation to the distribution L(t_n) of Student's statistic t_n does not hold uniformly over the class P_n of samples from zero-mean unit-variance bounded distributions. We present lower bounds to the corresponding error. We suggest a generalisation of the T-test that allows for variability of possible approximating distributions L(t_n).

14:00 Robust Inference On Infinite And Growing Dimensional Time Series Regression
Abhimanyu Gupta, Myung Hwan Seo

Abstract: We develop a class of tests for a growing number of restrictions in infinite and increasing order time series models such as multiple regression with growing dimension, infinite-order autoregression and nonparametric sieve regression. Examples include the Chow test, Andrews and Ploberger (1994)-type exponential tests, and testing of general linear restrictions of growing rank p. Notably, our tests introduce a new scale correction to the conventional quadratic forms that are recentered and normalized to account for diverging p. This correction accounts for a high-order long-run variance that emerges as p grows with sample size. We propose a bias correction via a null-imposed bootstrap to control finite sample bias without sacrificing power unduly. A simulation study stresses the importance of robustifying testing procedures against the high-order long-run variance even when p is moderate. The tests are illustrated with an application to the oil regressions in Hamilton (2003).

14:30 Debiased Semiparametric U-Statistics With an Application to Inequality of Opportunity
Juan Carlos Escanciano

Abstract: We construct locally robust/orthogonal moment functions in a semiparametric U-statistics setting. We use orthogonal moments to propose new debiased estimators and inferences in a variety of applications ranging from inequality of opportunity to distributional treatment effects. U-statistics with high dimensional estimated parameters arise naturally in these applications. We introduce a novel U-moment representation of the First Step Influence Function (U-FSIF) to take into account the effect of the first step estimation on an identifying quadratic moment. Adding the U-FISF to the identifying quadratic moment function gives rise to an orthogonal quadratic moment. Our leading application is to measuring inequality of opportunity, for which we provide a novel and simple debiased estimator, and the first available inferential methods. Other applications include the bipartite ranking problem and quadratic functionals of the distributions of potential outcomes. All these examples are characterized by the use of U-statistics, and do not have valid inferential methods available and/or require Donsker conditions which are not met in practice. We overcome these problems in a general framework by using orthogonal moment functions and cross-fitting for U-statistics. We give general and simple regularity conditions for asymptotic theory, and illustrate an improved finite sample performance in simulations for our debiased measures of inequality of opportunity. In an empirical application, we find that 46% of income inequality in Spain can be explained by circumstances.

15:00 Pearson Tests for Conditional Distributions.
Miguel A Delgado, Julius Vainora

Abstract: We propose Pearson type goodness-of-fit tests to check parametric continuous conditional distribution model specifications using grouped data. Observations of the Rosenblatt transform of the dependent variable and of the explanatory variables are simultaneously cross-classified and the corresponding joint frequencies are represented by a contingency table. As in the classical case for marginal distributions, the Pearson statistic is identical to the Wald and Lagrange multiplier statistics based on the grouped data likelihood, which is in turn asymptotically equivalent to the likelihood ratio statistic under the null hypothesis. We also provide a Chernoff-Lehmann result for the Pearson statistic using the raw data maximum likelihood estimator, which forms a basis to show that the corresponding Wald statistic is asymptotically distributed as a chi-squared with degrees of freedom invariant to the number of parameters in the model. The asymptotic distribution of the statistics do not change when the explanatory variables classification is data dependent. The finite sample properties of the tests are examined by means of Monte Carlo experiments.

Modern challenges in structured time series data

Organiser: YiYu

Chair: Yi Yu

Room: Akamas C

13:30 Adversarially robust change point detection
Mengchu Li, Yi Yu

Abstract: Change point detection is becoming increasingly popular in many application areas. On one hand, most of the theoretically-justified methods are investigated in an ideal setting without model violations, or merely robust against identical heavy-tailed noise distribution across time and/or against isolate outliers; on the other hand, we are aware that there have been exponentially growing attacks from adversaries, who may pose systematic contamination on data to purposely create spurious change points or disguise true change points. In light of the timely need for a change point detection method that is robust against adversaries, we study the simplest univariate mean change point detection problem. The adversarial attacks are formulated through the Huber ε-contamination framework, which allows the contamination distributions to be different at each time point. We demonstrate a phase transition phenomenon in change point detection and derive the minimax-rate optimal localisation error rate, quantifying the cost of accuracy in terms of the contamination proportion. We propose a computationally feasible method, matching the minimax lower bound under certain conditions, saving for logarithmic factors.

14:00 Network Estimation by Mixing: Adaptivity and More
Can Le, Tianxi Li

Abstract: Networks analysis has been commonly used to study the interactions between units of complex systems. One problem of particular interest is learning the network's underlying connection pattern given a single and noisy instantiation. While many methods have been proposed to address this problem in recent years, they usually assume that the true model belongs to a known class, which is not verifiable in most real-world applications. Consequently, network modeling based on these methods either suffers from model misspecification or relies on additional model selection procedures that are not well understood in theory and can potentially be unstable in practice. To address this difficulty, we propose a mixing strategy that leverages available arbitrary models to improve their individual performances. The proposed method is computationally efficient and almost tuning-free; thus, it can be used as an off-the-shelf method for network modeling. We show that the proposed method performs equally well as the oracle estimate when the true model is included as individual candidates. More importantly, the method remains robust and outperforms all current estimates even when the models are misspecified. Extensive simulation examples are used to verify the advantage of the proposed mixing method. Evaluation of link prediction performance on 385 real-world networks from six domains also demonstrates the universal competitiveness of the mixing method across multiple domains.

14:30 Random Subspace Ensemble
Yang Feng

Abstract: We propose a flexible ensemble framework, Random Subspace Ensemble (RaSE). In the RaSE algorithm, we aggregate many weak learners, where each weak learner is trained in a subspace optimally selected from a collection of random subspaces using a base method. In addition, we show that the number of random subspaces needs to be very large in a high-dimensional framework to guarantee that a subspace covering signals is selected. Therefore, we propose an iterative version of the RaSE algorithm and prove that under some specific conditions, a smaller number of generated random subspaces are needed to find a desirable subspace through iteration. We study the RaSE framework for classification, where a general upper bound for the misclassification rate was derived, and for screening, where the sure screening property was established. An extension called Super RaSE was proposed to allow the algorithm to select the optimal pair of base method and subspace during the ensemble process. The RaSE framework is implemented in the R package RaSEn on CRAN.

15:00 Optimal partition recovery: from chain graphs to lattices then general graphs.
Yi Yu

Abstract: In this talk, I will talk about the partition recovery problems, starting from chain graphs, which can be seen as 1-dimensional grid graphs, to general d-dimensional grid graphs and finally to general graphs. I will talk about the fundamental limits in finding a consistent partition in all these problems, together with polynomial-time rate-optimal methods.

Nonparametric random coefficient regression

Organiser: Hajo Holzmann

Chair: Hajo Holzmann

Room: Aphrodite A

13:30 Rate-optimal nonparametric estimation for random coefficient regression models Hajo Holzmann, Alexander Meister

Abstract: Random coefficient regression models are a popular tool for analyzing unobserved heterogeneity, and have seen renewed interest in the recent econometric literature. In this paper we obtain the optimal pointwise convergence rate for estimating the density in the linear random coefficient model over Hölder smoothness classes, and in particular show how the tail behavior of the design density impacts this rate. In contrast to previous suggestions, the estimator that we propose and that achieves the optimal convergence rate does not require dividing by a nonparametric density estimate. The optimal choice of the tuning parameters in the estimator depends on the tail parameter of the design density and on the smoothness level of the Hölder class, and we also study adaptive estimation with respect to both parameters.

14:00 Random coefficients when regressors have limited variation
Christophe Gaillac, *Eric Gautier*

Abstract: We consider a linear model where the coefficients - intercept and slopes - are random with a distribution in a nonparametric class and independent from the regressors. The main drawback of this model is that identification usually requires the regressors to have a support which is the whole space. This is rarely satisfied in practice. Rather, in this paper, the regressors can have a support which is a proper subset. This is possible by assuming that the slopes do not have heavy tails. Lower bounds on the supremum risk for the estimation of the joint density of the random coefficients density are derived for this model and a related white noise model. We present an estimator, its rates of convergence, and a data-driven rule which delivers adaptive estimators. The corresponding R package is https://CRAN.R-project.org/package=RandomCoefficients

14:30 Tests for Qualitative Features in the Random Coefficients Model
Katharina Proksch, Fabian Dunker, Konstantin Eckle, Johannes Schmidt-Hieber

Abstract: The random coefficients model is an extension of the linear regression model that allows for unobserved heterogeneity in the population by modeling the regression coefficients as random variables. Given data from this model, the statistical challenge is to recover in-formation about the joint density of the random coefficients which is a multivariate and ill-posed problem. Because of the curse of dimensionality and the ill-posedness, pointwise nonparametric estimation of the joint density is difficult and suffers from slow convergence rates. Larger features, such as an increase of the density along some direction or a well-accentuated mode can, however, be much easier detected from data by means of statistical tests. In this talk, we follow this strategy and construct tests and confidence statements for qualitative features of the joint density, such as increases, decreases and modes. We propose a multiple testing approach based on aggregating single tests which are designed to extract shape information on fixed scales and directions. Using recent tools for Gaussian approximations of multivariate empirical processes, we derive expressions for the critical value. We evaluate our method on simulated and real data.

15:00 Varying Random Coefficient Models
Christoph Breunig

Abstract: This paper analyzes unobserved heterogeneity when observed charac- teristics are modeled nonlinearly. The proposed model builds on varying random coefficients (VRC) that are determined by nonlinear functions of observed regressors and additively separable unobservables. This paper proposes a novel estimator of the VRC density based on weighted sieve minimum distance. The main example of sieve bases are Hermite func-tions which yield a numerically stable estimation procedure. This paper shows inference results that go beyond what has been shown in ordinary RC models. We provide in each case rates of convergence and also es-tablish pointwise limit theory of linear functionals, where a prominent example is the density of potential outcomes. In addition, a multiplier bootstrap procedure is proposed to construct uniform confidence bands. A Monte Carlo study examines finite sample properties of the estimator and shows that it performs well even when the regressors associated to RC are far from being heavy tailed. Finally, the methodology is applied to analyze heterogeneity in income elasticity of demand for housing.

New developments of semi- and nonparametric methods in biostatistics

Organiser: RonghuiXu
Chair: Ronghui Xu
Room: Aphrodite B

13:30 Statistical modelling of COVID-19 data: Putting generalised additive models to work
Ursula Berger

Abstract: During the COVID-19 pandemic, Generalised Additive Models (GAMs) have proven in numerous aspects to provide a successful tool allowing to obtain important data-driven insights on the COVID-19 disease. In this talk, we demonstrate the flexibility of GAMs in two pandemic-related applications. In the first application, we use GAMS to investigate in what manner infections in specific age groups are associated to those of other age groups. We particularly focus on the role of school pupils, who have been claimed to be the driving force in the spread of the disease. In this context, we demonstrate that in the GAM framework parameter estimates are independent of the (unknown) case detection rate, which plays a relevant role in COVID-19 surveillance data. Second, we model the count of hospitalisations and fatal infections, which are both only available with a temporal delay. We illustrate how correcting for this reporting delay through a nowcasting procedure can naturally be incorporated in a GAM framework as an offset term. The talk aims to demonstrate the practical and “off-the-shelf” usability of GAMs, putting in to work in relevant real world application.

14:00 Doubly Robust Estimation under the Marginal Structural Cox Model for a Binary Treatment
Denise Rava, Ronghui Xu, Jelena Bradic

Abstract: The marginal structural Cox model (MSM) has been widely used to draw causal inference from observational studies with survival outcomes. The typical estimation approach under the Cox MSM is inverse probability weighting (IPW) using a propensity score (PS) model, which is known to be inconsistent if the PS model is misspecified. Effort to protect against such model misspecification involves augmentation, which has been a challenge in the past due to the non-collapsibility of the Cox regression model. We develop an augmented IPW (AIPW) estimator with doubly robust (DR) properties including rate DR, that enables us to use machine learning and a large class of nonparametric methods, to overcome the non-collapsibility challenge. We study both the theoretical and empirical performance of the AIPW estimator, and apply it to the data from a cohort of Japanese men in Hawaii followed since 1960s in order to study the effect of mid-life alcohol exposure on late-life mortality.

14:30 Semiparametric Estimation for Non-randomly Truncated Data
Ronghui Xu, Yuyao Wang, Andrew Ying

Abstract: In prospective cohort studies when patients enroll sequentially, time-to-event outcome can be subject to left truncation when only subjects with event time larger than enrollment time are included. In such cases, subjects with early event times tend not to be captured, leading to selection bias and invalidating the analysis results if simply ignoring the left truncation. Conventional methods adjusting for left truncation typically hinges heavily on the ``random truncation'' assumption that the truncation time and the event time are independent (conditional on covariates if using a regression model), which is usually subject to violation in practice. When the truncation time depends on additional covariates, inverse probability of truncation weighting can be used to adjust for left truncation. These, however, are sensitive to model misspecification. This motivates us to seek doubly robust estimators that provide extra protection against model misspecification. To that end, we first leverage the semiparametric theory to find the efficient score of any transformed survival time in the presence of non-random left truncation. We then use the efficient score to construct estimators that are shown to enjoy model double-robustness, that is, they are consistent and asymptotically normal (CAN) when one of the two nuisance parameters is consistently estimated at root-n rate, but not necessarily both; and rate double-robustness, that is, they are CAN when both of the nuisance parameters are consistently estimated and the product of the convergence rates of the two nuisance models is faster than root-n. The proposed doubly robust estimators can be easily extended to estimate the average treatment effect for randomized trials in the presence of non-random left truncation. Simulation studies illustrate the finite sample performance of our estimators. We apply our estimators to an Alzheimer's disease data set.

OODA: Trees and Graph Structured Objects

Organiser: Steve Marron

Chair: Steve Marron

Room: Christian Barnard

13:30 Regression Model with Unlabelled Network Response on Graph Space.
Anna Calissano, Aasa Feragen, Simone Vantini

Abstract: Graphs are a useful mathematical representation for different phenomena in different application fields, such as chemistry, medicine, transportation, and so-cial science. The analysis of populations of unlabelled networks is thus a promising but challenging task. In this work, unlabelled networks with Euclidean attributes are described in Graph Space, where every equivalence class represents all the networks obtained by permuting nodes. We hereby introduce a Generalized Geodesic Regression to model scalar-on-network relationships. Generalized Geodesic Regression is computed using the Align All and Compute Algorithm. Two case studies are described to showcase the potential of the model: the effect of the lockdown in the usage of public transport network in Copenhagen; the player passing network as a function of the match outcome during Fifa 2018 Championship.

14:00 Manifold valued data analysis of samples of Networks
Katie Severn, Ian Dryden, Simon Preston

Abstract: Networks can be used to represent many systems such as text documents and brain activity, and it is of interest to develop statistical techniques to compare networks. A general framework is developed for extrinsic statistical analysis of samples of networks, motivated by networks representing text documents in corpus linguistics. Networks are identified by their graph Laplacian matrices, for which metrics, embeddings, tangent spaces, and a projection from Euclidean space to the space of graph Laplacians are defined. This framework provides a way of computing means, performing principal component analysis, regression, and performing hypothesis tests, such as for testing for equality of means between two samples of networks. We apply the methodology to the set of novels by Jane Austen and Charles Dickens showing how our methods can distinguish differences between the two authors writing style and how each author's writings changed with time.

14:30 Null hypothesis testing for network-valued data with a multiscale approach: analysis of brain networks of patients with autism.
Alessia Pini, Ilenia Lovato, Stamm Aymeric, Maxime Taquet, Simone Vantini

Abstract: Networks have been extensively used for representing the human brain and for studying its structure and functions. The human brain is represented as a network whose knots are different location in the cortex and whose edges are the connections between such locations. In this framework, the investigation of neurological disorders can be performed by comparing samples of brain networks between healthy and non-healthy subjects, or between subjects affected by different types of disorders. This problem falls under the object-oriented data analysis paradygm: for each unit, the object of the statistical analysis is not a number nor a vector, but a complex object that is in this case a whole network representing brain connections. The aim of this work is to develop a method to test for differences between two populations of networks. The first target is to determine whether the two samples of networks have been generated by two different distributions. Then, it is key - for understanding the biological mechanisms involved in the pathology - to localize the differences on the brain network. This can be done by localizing the connections that present statistically significant differences between healthy subjects and patients. Motivated by this question, we propose a non-parametric finite-sample exact statistical test that allows to test for differences within and between pre-specified regions of interest of the network. The test is applied to brain structural connectivity networks obtained from electroencephalography, to differentiate children with non-syndromic autism from children with both autism and tuberous sclerosis complex.

15:00 Finite sample smeariness of Fréchet means on manifolds
Stephan Huckemann, Benjamin Eltzner, Shayan Hundrieser

Abstract: The seminal central limit theorem on manifolds by Bhattacharya and Patrangenaru (2005) asserts that under rather general assumptions, the limiting fluctuation of intrinsic Fréchet sample means in a local chart behaves just as its Euclidean kin. Recently it has been found that there are exceptional cases, exhibiting slower rates, leading to the more general smeary central limit theorem. While these exceptional cases are rather rare, not so rare is their effect on distributions of test statistics (e.g. for two-sample Hotelling tests) in the finite sample regime, often making quantile based tests inapplicable. Suitably designed bootstrap tests, however, remain valid. We make the underlying new concept of finite sample smeariness (FSS) precise and illustrate it. This is joint work with Benjamin Eltzner and Shayan Hundrieser

Data Analysis on Stratified Spaces with applications to SARSCov2 RNA analysis

Organiser: Vic Patrangenaru
Chair: Vic Patrangenaru
Room: Leda

13:30 Wald Space for Statistics of Phylogenetic Trees
Jonas Lueg, *Stephan F. Huckemann*, Tom M. W. Nye, Maryam K. Garba

Abstract: In order to conduct statistical inference on phylogenetic trees, it is necessary to equip the space of phylogenetic trees with a suitable geometry. We propose a new geometry that, in biological motivation, is based on an approximation of a simple statistical mutation model. It is induced by the restriction of the information geometry on the space of symmetric positive definite matrices to the tree-like matrices. This structure appears biologically and statistically more meaningful than the classical geometry proposed by Billera Holmes and Vogtman (BHV), 2001. Should not two phylogenetic trees with mutation distances diverging to infinity become similar and tend to the totally disconnected forest? This is so in our new “wald space” that encompasses phylogenetic forests modeling (some) evolutionary independence as limit cases. Our space also seems to feature less “stickiness” than BHV space: Stickiness describes the phenomenon that Fréchet means of samples with varying tree topologies can be constant beyond a random but finite sample size, a dead end for asymptotic statistics. Further, we show that our wald space is a geodesic, Riemann stratified space, satisfying Whitney condition A. We also propose an algorithm to compute geodesics, laying foundations for statistical analysis on wald space. This is joint work with Jonas Lueg, Tom Nye and Maryam Garba

14:00 Investigating Two Possible Origins of SARS-CoV-2: an RNA Analysis on Tree Spaces,
Roland Moore, Victor Patrangenaru, *Adam Dixon*

Abstract: Using the construction of Billera, Holmes, and Vogtmann, rooted phylogenetic trees with three leaves (3-trees) are regarded as a two-dimensional stratified space. Since 3-trees include exactly one interior node and one interior edge, branch lengths corresponding to the root-to-interior-node edge and the interior edge are used to define an evolutionary distance between the root and the most recent common ancestor (MRCA) of the leaves. With this definition, the proximities of various roots to the leaves can be compared. To apply this method, recent RNA sequences of SARS-CoV-2 collected from multiple sources are aligned and used to compare sets of 3-trees with two different roots: bat coronavirus RatG13 (BatCov RatG13) and a SARS-CoV-2 sequence taken from a human subject at the onset of the COVID-19 pandemic. Nonparametric bootstrap methods are used to infer the difference in mean evolutionary distance between the two root groups, and the early-sequenced SARS-CoV-2 root appears to be evolutionarily closer to the MRCA than BatCov RaTG13.

14:30 CLT on Stratified Spaces with an example of phylogenies of SARS-CoV-2 data
Vic Patrangenaru, Chen Shen

Abstract: Important additions to the MSC2020 were made with the introduction of 62R20 - Statistics on metric spaces and 62R30 Statistics on manifolds. From the perspective of nonparametric estimation of location, the complete separable metric structure on an object space leads to consistency of Frechet sample means of distributions to their population counterparts. This general metric structure is nevertheless insufficient for proving asymptotic results for such indices. An richer metric structure, allowing for some level of smoothness to prove the CLT was needed. Such a structure on an object space, first introduced at CRM 2011 (see Bhattacharya et al(2013)[1]) is that of stratified space. Special cases of interest are BHV spaces, and rooted tree spaces (Moore et al(2021)[2], which in case of trees with three leafs are open books (Hotz et al(2013)[4]). Here, for simplicity we analyze the behavior of intrinsic means of distributions of SARS-CoV-2 RNA sequences from various parts of the World (Shen(2021)[3]).

Statistical Inference in Models with Functional Data

Organiser: Wenceslao Gonzalez Manteiga
Chair: Pedro Galeano
Room: Athena

13:30 Bootstrap test for the equality of distributions in separable Hilbert spaces
Ana Colubi, Gil Gonzalez-Rodriguez, Wenceslao Gonzalez-Manteiga, Manuel Febrero-Bande

Abstract: A test with bootstrap calibration is developed to check if two independent random elements taking values in a separable Hilbert space have the same distribution. The test is based on a so-called energy statistic. This statistic can be expressed alternatively as a comparison of means of a transformation of the random elements into a new separable Hilbert space. In this way, the equality of distributions is equivalent to the equality of expectations for the transformed random elements, and consequently, a bootstrap test of equality of means can be employed to solve the equality of distributions test. The test can be applied with simple operations in the original space, and the bootstrap approaches, namely exchangeable weighted bootstrap and wild bootstrap, are shown to be asymptotically correct and consistent.

14:00 On the complexity index of a functional time series
Kwo Lik (Lax) Chan, Aldo Goia, Enea Bongiorno

Abstract: Consider a functional time series taking values in a general topological space and assume that its Small-Ball Probability (SmBP) factorizes into two terms that play the role of a surrogate density and of a volume term. The latter is a mean for studying the complexity of the underlying process, as the volume term may reveal some latent feature of the process. In some cases, it can be analytically specified in a parametric form: this current work focuses on situations when the processes belong to the monomial family, like in the finite dimensional and fractal case, for which the volume term has a monomial form depending on the SmBP radius and a parameter named complexity index. This work presents some results concerning the study of a nonparametric estimator for such complexity index in the beta-mixing framework. As a by-product, a nonparametric estimator for the volume term based on a U-statistic is also investigated. Weak consistency of these estimators are provided. In the particular case of a monomial family, it is possible to estimate the complexity index by minimizing a suitable dissimilarity measure. For this estimator asymptotic Gaussianity is shown, providing a theoretical justification to build confidence interval for the complexity index. A Monte Carlo simulation is carried out in order to assess the performance of the methodology for finite sample sizes. Finally, the new method is applied to detect the complexity of a real world dataset.

14:30 A goodness-of-fit test for functional time series with applications to diffusion processes
Javier Álvarez-Liébana, Wenceslao González-Manteiga, Alejandra López-Pérez, Manolo Febrero- Bande

Abstract: Within the burgeoning Functional Data Analysis framework, the analysis of intra-day high-frequency data is currently one of the topics of greatest interest in financial research. In this context, the Functional Linear Model with Functional Response is one of the most relevant models to assess the relation between two functional random variables. A particular case arises when functional responses are given by their own past values, in which functional errors and responses are (linearly) correlated. In this work, a novel goodness-of-fit test for autoregressive Hilbertian (ARH) models is proposed. The test imposes no restrictions on the functional form of the linear regression operator, and the test statistic is formulated in terms of a Cramér–von Mises norm. A wild bootstrap resampling procedure is used for calibration, such that the finite sample behavior of the test, regarding power and size, is checked via a simulation study. Furthermore, we also provide a new specification test for stochastic diffusion models, such as Ornstein-Uhlenbeck processes, illustrated with an application to intra-day currency exchange rates. In particular, a two-stage methodology is proffered: firstly, we check if functional samples and their past values are related via ARH(1) model; secondly, under linearity, we perform a functional F-test.

15:00 Goodness-of-fit tests for the functional linear model with scalar response with responses missing at random
Manuel Febrero-Bande, *Pedro Galeano*, Eduardo García-Portugués, Wenceslao González-Manteiga

Abstract: We construct goodness-of-fit tests for the Functional Linear Model with Scalar Response (FLMSR) when some of the responses are Missing At Random (MAR). For that, we extend two testing procedures recently proposed for the case in which all the responses are observed to the MAR case. The test statistics associated to the two testing procedures are built from marked empirical processes indexed by the randomly projected functional covariate and depend on proper estimates of the functional slope of the FLMSR. They are relatively easy to compute and their distributions under the null hypothesis are simple to calibrate based on wild bootstrap procedures. The behaviour of the two statistics in conjunction with different estimates of the functional slope of the model when some of the responses are MAR are compared by means of an extensive simulation study. Additionally, the testing procedures are applied to two real data sets for checking whether the linear hypothesis holds.

15:30 - 16:00 Coffee Break

16:00 - 18:00 Invited Paper Session 8

Nonparametric approaches in survival analysis
Organiser: Ingrid Van Keilegom
Chair: Ingrid Van Keilegom
Room: Akamas A

16:00 Instrumental variable estimation of dynamic treatment effects on a survival outcome
Jad Beyhum, Samuele Centorrino, Jean-Pierre Florens, Ingrid Van Keilegom

Abstract: This paper considers identification and estimation of the causal effect of the time Z until a subject is treated on a survival outcome T. The treatment is not randomly assigned, T is randomly right censored by a random variable C and the time to treatment Z is right censored by min(T,C). The endogeneity issue is treated using an instrumental variable explaining Z and independent of the error term of the model. We study identification in a fully nonparametric framework. We show that our specification generates an integral equation, of which the regression function of interest is a solution. We provide identification conditions that rely on this identification equation. For estimation purposes, we assume that the regression function follows a parametric model. We propose an estimation procedure and give conditions under which the estimator is asymptotically normal. The estimators exhibit good finite sample properties in simulations. Our methodology is applied to find evidence supporting the efficacy of a therapy for burn-out.

16:30 A new lack-of-fit test for censored quantile regression models when the covariate is high-dimensional
Mercedes Conde-Amboage, Ingrid Van Keilegom, Wenceslao González-Manteiga

Abstract: Quantile regression was introduced by Koenker and Basset (1978, Econometrica) as a weighted absolute residuals fit which allows to extend some properties of classical least squares estimation to quantile regression estimates. This kind of regression allows a more detailed description of the behaviour of the response variable, adapts to situations under more general conditions of the error distribution (that is, do not require stringent assumptions, such as homoscedasticity or normality) and enjoys properties of robustness. Hereby it facilitates a more complete and robust analysis of the information. For all that, quantile regression is a very useful statistical technology for a large diversity of disciplines. In particular, quantile regression provides good results when complex data are considered, for instance, when the response variable is right-censored. Along this talk, a new lack-of-fit test for censored quantile regression models with multiple covariates will be presented. The test is based on the cumulative sum of residuals with respect to unidimensional linear projections of the covariates. The test is then adapting the ideas of Escanciano (2006, Econometric Theory) to cope with high-dimensional covariates, to the test proposed by Conde-Amboage et al (2021, Scandinavian Journal of Statistics). It will be shown the limit distribution of the empirical process associated with the test statistic. Furthermore, in order to approximate the critical values of the test, a wild bootstrap mechanism is used, which is similar to that proposed by Orbe and Núnez-Antón (2013, Communications in Statistics-Simulation and Computation). In addition, an extensive simulation study and an interesting application to real data will be presented in order to show the behaviour of the new test in practice.

17:00 Causal inference for semi-competing risks data
Daniel Nevo, Malka Gorfine

Abstract: An emerging challenge for time-to-event data is studying semi-competing risks, namely when two event times are of interest: a non-terminal event (e.g. Alzheimer's disease diagnosis) time, and a terminal event (e.g. death) time. The non-terminal event is observed only if it precedes the terminal event, which may occur before or after the non-terminal event. Studying treatment or intervention effects on the dual event times is complicated because for some units, the non-terminal event time may occur under one treatment value but not under the other. Until recently, existing approaches generally disregarded the time-to-event nature of both outcomes. More recent research focused on principal strata effects within time-varying populations coupled with Bayesian estimation. In this paper, we present alternative estimands, based on a single stratification of the population, corresponding to the scientific questions of interest. We present a novel assumption utilizing the time-to-event nature of the data, that is generally more flexible than the often-invoked monotonicity assumption. Our new assumption enables partial identifiability of causal effects of interest. We further discuss a frailty-based sensitivity analysis approach, and give conditions under which full identification is possible. We present non-parametric and semi-parametric estimation methods under right censoring. We illustrate the utility of our approach in a study of the causal effects of having APOE e4 allele on late-onset Alzheimer's disease and death.

Bayesian Nonparametrics: theory and computation

Organiser: Sergios Agapiou

Chair: Sergios Agapiou

Room: Akamas C

16:00 Complexity of coordinate-wise inference algorithms for crossed random effects
Omiros Papaspiliopoulos, Giacomo Zanella, Swanardip Ghosh

Abstract: We analyze the computational complexity of coordinate-wise algorithms, such as backfitting, coordinate ascent variational inference and Gibbs sampling for high-dimensional crossed random effect models with several categorical factors. We establish that certain implementations of such algorithms are scalable, in the sense that the overall complexity (time per iteration multiplied by the required number of iterations) scales linearly in the number of parameters and data. In fact under certain assumptions such algorithms enjoy a blessing of dimensionality whereby their convergence rate improves with dimension. The theory leverages on state of the art results for random walks on random bi-regular graphs and the so-called Alon's conjecture.

16:30 A Bayesian Nonparametrics Analysis of Finite Element and Graphical Representations of Gaussian Processes
Daniel Sanz-Alonso, Ruiyi Yang

Abstract: Gaussian processes (GPs) are popular models for random functions in computational and applied mathematics, statistics, machine learning, and data science. However, GP methodology scales poorly to large data-sets due to the need to factorize a dense covariance matrix. In spatial statistics, a standard approach to surmount this challenge is to represent Matérn GPs using finite elements, obtaining an approximation with a sparse precision matrix. The first part of the talk will give new understanding of this approach for regression and classification with large data-sets, showing that under mild smoothness assumptions the dimension of the matrices that need to be factorized can be reduced without hindering the estimation accuracy. The analysis balances finite element and statistical errors to show that there is a threshold beyond which further refining of the discretization increases the computational cost without improving the estimation accuracy. In the second part of the talk, I will introduce graphical representations of GPs to model random functions on high-dimensional point clouds, greatly expanding the important but limited scope of the finite element approach. I will show error bounds on the graphical representations, and study the associated posterior contraction in a semi-supervised learning problem. Time permitting, I will demonstrate the versatility of the graphical approach in applications to regression, classification and PDE-constrained Bayesian inverse problems where the covariates are sampled from a variety of hidden manifolds.

17:00 A unified construction for series representations and finite approximations of completely random measures.
Xenia Miscouridou

Abstract: Infinite-activity completely random measures (CRMs) have become important building blocks of complex Bayesian nonparametric models. They have been successfully used in various applications such as clustering, density estimation, latent feature models, survival analysis or network science. Popular infinite-activity CRMs include the (generalized) gamma process and the (stable) beta process. However, except in some specific cases, exact simulation or scalable inference with these models is challenging and finite-dimensional approximations are often considered. In this work, we propose a general and unified framework to derive both series representations and finite-dimensional approximations of CRMs. Our framework can be seen as an extension of constructions based on size-biased sampling of Poisson point process (Perman et al. 1992). It includes as special cases several known series representations as well as novel ones. In particular, we show that one can get novel series representations for the generalized gamma process and the stable beta process. We also provide some analysis of the truncation error.

17:30 Linear methods for non-linear inverse problems
Botond Szabo, Aad van der Vaart, Geerten Koers

Abstract: We consider recovering an unknown function f from a noisy observation of the solution u to a partial differential equation, where for the elliptic differential operator L, the map L(u) can be written as a function of u and f, under Dirichlet boundary condition. A particular example is the time-independent Schrödinger equation. We transform this problem into the linear inverse problem of recovering L(u), and show that Bayesian methods for this problem may yield optimal recovery rates not only for u, but also for f. The prior distribution may be placed on u or its elliptic operator. Adaptive priors are shown to yield adaptive contraction rates for f, thus eliminating the need to know the smoothness of this function. Known results on uncertainty quantification for the linear problem transfer to f as well. The results are illustrated by several numerical simulations.

Recent advances in nonparametric and semiparametric statistics

Organiser: LanWang

Chair: Irène Gijbels

Room: Aphrodite A

16:00 Clustering on the Sphere: State-of-the-art and a Poisson kernel-based method
Marianthi Markatou

Abstract: Many applications of interest involve data that can be analyzed as unit vectors on a d-dimensional sphere. Specific examples include text mining, biology, astronomy and medicine. We present a clustering method based on mixtures of Poisson-kernel based densities on the high-dimensional sphere. We study connections of the Poisson kernel-based densities with other distributions appropriate for the analysis of directional data, prove identifiability of mixtures of the Poisson kernel-based densities model, and discuss convergence of the associated EM-type algorithm. Furthermore, we propose a method to simulate data from Poisson kernel-based densities and discuss in detail implementation aspects of the algorithm such as initialization and stopping rules. We then exemplify our methods via application on real data sets and simulation experiments. Our experimental results show that the newly introduced model exhibits higher macro-precision and macro-recall than competing methods based on von Mises Fisher and Watson distributions. This is joint work with Mojgan Golzy, Ph.D.

16:30 "Flexible Semi-Parametric Gaussian Process Classification".
Dipak Dey, Zhiyong Hu

Abstract: Predicting binary response data is one of the most commonly encountered problems in the field of statistics. The simplest and the most frequently used model for such binary classification job is the logistic regression model. However, the systematic component of logistic regression is often limited to a linear or parametric form, which restricts the model's flexibility in fitting the data. Moreover, the popular link functions, such as probit, logit, and complementary log-log links have constant skewness that cannot learn the shape of link from the data, which in turn limits the flexibility of modeling again. Thus, in this work we propose a flexible semi-parametric Gaussian process classification model that incorporates skewed Weibull link. By placing a non-parametric Gaussian process prior over the latent systematic component and utilizing the flexible Weibull link function that can approximate the commonly-adopted link functions, great flexibility in modeling can be achieved. The performance of the proposed model is demonstrated in both simulation experiment and real data analyses, by comparing with several alternative models. The sparse version of the proposed model which has O(nm2) complexity is also implemented to accommodate the big data scenario, where traditional Gaussian process model is restricted by its prohibitive O(n 3) complexity.

17:00 Gaussian copulas adjusted for non- and semi-parametric regression
Irène Gijbels, Yue Zhao, Ingrid Van Keilegom

Abstract: We consider a multivariate response regression model where each coordinate is described by a location-scale non- or semi-parametric regression model. The dependence in the error terms follows a Gaussian copula, and the primary interest is in the estimation of the associated copula correlation matrix given a sample of the response and the covariate. We consider a normal scores estimator based on residual ranks calculated from preliminary non- or semi-parametric estimators of the location and scale functions. It is shown that the residual-based normal scores estimator is asymptotically equivalent to its oracle counterpart, and the explicit rate of convergence is established. The methodology is applied to several nonparametric and semiparametric location-scale regression settings.

Statistical Topological Data Analysis and Geometric Inference

Organiser: Johannes Krebs

Chair: Daniel Rademacher

Room: Aphrodite B

16:00 Goodness-of-fit tests for spatial tessellations based on the persistence diagram
Christian Hirsch, Johannes Krebs, Claudia Redenbach

Abstract: n applications in materials science, it is common to work with tessellation data where the cell centers are not scattered entirely at random but are subject to repulsive interactions. Therefore, Gibbs-Voronoi and Gibbs-Laguerre tessellations are important building blocks when constructing stochastic geometry models. Moreover, recently the persistence diagram has become a popular tool to detect subtle topological features in data. Based on the framework in (Schreiber & Yukich, 2013), I will present a functional CLT for the persistence diagram on Gibbsian tessellations, which can form the basis for goodness-of-fit tests. This talk is based on joint work with J. Krebs and C. Redenbach.

16:30 Large deviation principle for geometric and topological functionals and associated point processes
Takashi Owada, Christian Hirsch

Abstract: A large deviation principle is proved for the point process associated with k-element connected components in the d-dimensional Euclidean space with respect to the connectivity radii as a function of sample size. The random points are generated from a homogeneous Poisson point process so that the connectivity radius is of the so-called sparse regime. The rate function for the obtained large deviation principle can be represented as relative entropy. As an application, we deduce large deviation principles for various functionals and point processes appearing in stochastic geometry and topology. As concrete examples of topological invariants, we consider persistent Betti numbers of geometric complexes and the number of Morse critical points of the min-type distance function.

17:00 Statistical Query Complexity of Manifold Estimation
Eddie Aamari

Abstract: The statistical query (SQ) framework consists in replacing the usual access to samples from a distribution, by the access to adversarially perturbed expected values of functions interactively chosen by the learner. This framework provides a natural estimation complexity measure, enforces robustness through adversariality, and is closely related to differential privacy. In this talk, we study the SQ complexity of estimating d-dimensional submanifolds in R^n. We propose a purely geometric algorithm called Manifold Propagation, that reduces the problem to three local geometric routines: projection, tangent space estimation, and point detection. We then provide constructions of these geometric routines in the SQ framework. Given an adversarial STAT(τ) oracle and a target precision ε=Ω(τ^{2/(d+1)}) for the Hausdorff distance, the resulting SQ manifold reconstruction algorithm has query complexity O(npolylog(n)ε^{d/2}), which is proved to be nearly optimal. In the process, we will present low-rank matrix completion results for SQ's and lower bounds for (randomized) SQ estimators in general metric spaces.

17:30 Topological Inference via Bootstrap Resampling
Benjamin Roycraft

Abstract: Multivariate bootstrap procedures for general stabilizing statistics are considered, with specific application to topological data analysis. Existing limit theorems for topological statistics prove difficult to use in practice for the construction of confidence intervals, motivating the use of the bootstrap in this capacity. However, the standard nonparametric bootstrap does not directly provide for asymptotically valid confidence intervals in some situations. A smoothed bootstrap procedure, instead, is shown to give consistent estimation in these settings. Specific statistics considered include the persistent Betti numbers of Cech and Vietoris-Rips complexes over point sets in Euclidean space, along with Euler characteristics, and the total edge length of the k-nearest neighbor graph.

Extremes and Machine Learning

Organiser: Anna Kiriliouk

Chair: Ioannis Papastathopoulos

Room: Christian Barnard

16:00 Extreme value statistics in semi-supervised models
John Einmahl

Abstract: We consider extreme value analysis in a semi-supervised setting, where we observe, next to the n data on the target variable, n + m data on one or more covariates. This is called the semi-supervised model with n labeled and m unlabeled data. We derive an estimator for the extreme value index of the target variable in this setting and establish its asymptotic behavior. Our estimator improves the univariate estimator, based on only the n target variable data, in terms of asymptotic variance whereas the asymptotic bias remains unchanged. We present a simulation study in which the asymptotic results are confirmed and also an extreme quantile estimator is derived and shown to perform well. Finally the estimation method is applied to rainfall data in France. This is joint work with Hanan Ahmed and Chen Zhou.

16:30 Bayesian semi-parametric modeling of jointly heteroscedastic extremes
Karla Vianey Palacios Ramirez

Abstract: In this paper we introduce a statistical method for modeling the frequency of joint extremes over time. The joint scedasis function for bivariate extremes; here introduced as a function that carries information on the frequency of joint extremes over time. We develop Bayesian estimators for the two parameters of interest, the extreme value index and the scedasis function. Bayesian inference is proposed to estimate the scedasis function based on mixture of Polya trees.

17:00 Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models
Gilles Stupfler, Stéphane Girard, Antoine Usseglio-Carleve

Abstract: Expectiles define a least squares analogue of quantiles. They have been the focus of a substantial quantity of research in the context of actuarial and financial risk assessment over the last 10 years. Unlike quantiles, expectiles induce coherent risk measures and are calculated using tail expectations rather than merely tail probabilities; contrary to the popular quantile-based Expected Shortfall, they define elicitable risk measures. The behaviour and estimation of extreme expectiles using independent and identically distributed heavy-tailed observations has been investigated in a recent series of papers. The case of extreme conditional expectile estimation has, however, not been addressed so far in the literature. We build here a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise. We demonstrate how our results can be applied to a wide class of important examples, among which linear heteroscedastic models, heteroscedastic single-index models and autoregressive time series models. We showcase our estimators on finite-sample experiments.

Analysis of high-dimensional complex data

Organiser: Eugen Pircalabelu

Chair: Eugen Pircalabelu

Room: Leda

16:00 Sufficient Dimension Reduction for high-dimensional settings
Andreas Artemiou, Eugen Pircalabelu

Abstract: In this talk we will propose a new algorithm for Sufficient Dimension Reduction for high-dimensional data. Large p - small n settings are challenging in most Sufficient Dimension Reduction algorithms due to the need of using inverse covariance matrices. In this work, we expand a method called Principal Support Vector Machine (Li, Artemiou, Li, AOS 2011) to address high-dimensional settings. We are using principal projections of the non-invertible covariance matrix, to solve a problem in a lower dimensional space. First we demonstrate that the solutions of the the high-dimensional and the lower-dimensional problems are equivalent. Then we demonstrate that the solution to the lower-dimensional problem is equivalent to solving. We show that our method performs better under some settings to similar methods in the literature.

16:30 Stationary Subspace Analysis - A Statistical Perspective
Klaus Nordhausen, Lea Flumian, Markus Martilainen
Abstract: Multivariate times series occur in many application areas and are challenging to model. A common approach is therefore to assume that the observed time series can be decomposed into latent components with different exploitable properties. In some of these models especially nonstationary components are of interest and thus the nonstationary subspace should be separated from the stationay subspace which is often referred to as stationary subspace analysis (SSA). Different methods are considered here for this purpose and a test suggested to make inference about the dimensions of the subspaces.

17:00 Inference for the explained Gini coefficient: a penalized bootstrap procedure
Alexandre Jacquemain, Cédric Heuchenne, Eugen Pircalabelu

Abstract: The explained Gini coefficient, introduced by Heuchenne and Jacquemain (2022), is a measure of the economic inequality that can be attributed to a set of covariates. Similarly to the R^2 in regression models, this quantity never decreases as we keep introducing new covariates and may suffer from overfitting on datasets of large dimension. We propose a penalized bootstrap procedure which selects the relevant covariates and produces inference for the explained Gini coefficient, while avoiding overfitting. The obtained estimator achieves the Oracle property and can be computed efficiently. In this respect, we introduce the SCAD-FABS algorithm, an adaptation of the FABS algorithm proposed by Shi et al. (2018) to the SCAD penalty. The performance of the procedure is established by theoretical guarantees and assessed via Monte-Carlo simulations.

17:30 The Completion Of Covariance Kernels
Kartik Waghmare

Abstract: We consider the problem of positive-definite continuation: extending a partially specified covariance kernel from a subdomain Ω of a rectangular domain $I \times I$ to a covariance kernel on the entire domain $I \times I$. For abroad class of domains Ω called serrated domains, we are able to present a complete theory. Namely, we demonstrate that a canonical completion always exists and can be explicitly constructed. We characterise all possible completions as suitable perturbations of the canonical completion, and determine necessary and sufficient conditions for a unique completion to exist. We interpret the canonical completion via the graphical model structure it induces on the associated Gaussian process. Furthermore, we show how the estimation of the canonical completion reduces to the solution of a system of linear statistical inverse problems in the space of Hilbert-Schmidt operators, and derive rates of convergence under reasonable source conditions. We conclude by providing extensions of our theory to more general forms of domains, and by demonstrating how our results can be used to construct covariance estimators from sample path fragments of the associated stochastic process. Our results are illustrated numerically by way of a simulation study and a real example.

Topics in nonparametric inference and estimation

Organiser: Hira L. Koul & Indeewara Perera

Chair: Indeewara Perera

Room: Athena

16:00 Testing for Restricted Stochastic Dominance under Survey Nonresponse with Panel Data
Rami Tabri, Matthew Elias

Abstract: Stochastic dominance relations frequently arise across various areas of applied statistics for conducting distributional analyses. In practice, statistical tests are implemented to infer such relations using panel survey data, which are predicated on the assumption that the practitioner has access to a complete dataset. This assumption is typically violated in empirical practice as survey nonreponse is inevitable. This paper develops a testing procedure for restricted stochastic dominance with matched pair data from panel surveys under nonresponse, using the worst-case bounds on the distributions. The advantage of using these bounds in distributional comparisons is that conclusions are robust to the nature of the nonresponse-generating process. The testing procedure uses the method of pseudo-empirical likelihood to formulate the test statistic and compares it to a critical value from the chi-squared distribution with one degree of freedom. This paper develops the asymptotic properties of the testing procedure under the null and alternative hypotheses, establishing its asymptotic validity with uniformity and its performance against distant and local alternatives. Finally, we illustrate the test by analysing the distribution of household incomes in Australia using the HILDA survey

16:30 Seasonal Cyclical Models
Natalia Bailey, Karim Abadir, Walter Distaso, Liudas Giraitis

Abstract: A number of economic, financial and climatic time series are characterised by persistent periodic movements. Such stochastic cycles are distinguished by their dependence patterns characterised by memory parameters and peaks in the spectrum. For example, the seasonal Gaussian Gegenbauer process is stationary when its memory parameter is less than 1/2 and its spectral density has poles at frequencies away from zero. When memory parameter is greater than 1/2, the Gegenbauer process is considered to be non-stationary. In this paper, we develop asymptotic theory for an estimator that detects seasonal peaks in the periodogram of a given process. The precision of this detection is a function of the process's memory parameter. Confidence bands for the location of a pole in the spectrum range within one or two frequencies around the estimated peak as the time series approaches the non-stationary region. We conduct a detailed Monte Carlo simulation study that examines the small sample properties of our estimator and results are in line with the theoretical findings. As an empirical illustration, we use our estimator to examine whether well-established seasonal adjustment methods (e.g. the x11 specification of X-12-ARIMA) are successful in eliminating all traces of seasonality in selected economic time series.

17:00 Optimal bias correction of the log-periodogram estimator of the fractional parameter: A jackknife approach
Kanchana Nadarajah, Gael M Martin, Donald S Poskitt

Abstract: We use the jackknife to bias correct the log-periodogram regression (LPR) estimator of the fractional parameter in a stationary fractionally integrated model. The weights for the jackknife estimator are chosen in such a way that bias reduction is achieved without the usual increase in asymptotic variance, with the estimator viewed as 'optimal' in this sense. The theoretical results are valid under both the non-overlapping and moving-block sub-sampling schemes that can be used in the jackknife technique and do not require the assumption of Gaussianity for the data generating process. A Monte Carlo study explores the finite sample performance of different versions of the jackknife estimator, under a variety of scenarios. The simulation experiments reveal that when the weights are constructed using the parameter values of the true data generating process, a version of the optimal jackknife estimator almost always outperforms alternative semi-parametric bias-corrected estimators. A feasible version of the jackknife estimator, in which the weights are constructed using estimates of the unknown parameters, whilst not dominant overall, is still the least biased estimator in some cases. Even when misspecified short-run dynamics are assumed in the construction of the weights, the feasible jackknife estimator still shows a significant reduction in bias under certain designs. As is not surprising, parametric maximum likelihood estimation outperforms all semi-parametric methods when the true values of the short memory parameters are known but is dominated by the semi-parametric methods (in terms of bias) when the short memory parameters need to be estimated, in particular when the model is misspecified.

17:30 Specification tests for GARCH processes with nuisance parameters on the boundary
Indeewara Perera

Abstract: This paper develops tests for the correct specification of the conditional variance function in GARCH models when the true parameter may lie on the boundary of the parameter space. The test statistics considered are of Kolmogorov-Smirnov and Cramer-von Mises type, and are based on a certain empirical process marked by centered squared residuals. The limiting distributions of the test statistics depend on unknown nuisance parameters in a non-trivial way, making the tests difficult to implement. We therefore introduce a novel bootstrap procedure which is shown to be asymptotically valid under general conditions, irrespective of the presence of nuisance parameters on the boundary. The proposed bootstrap approach is based on shrinking of the parameter estimates used to generate the bootstrap sample toward the boundary of the parameter space at a proper rate. It is simple to implement and fast in applications, as the associated test statistics have simple closed form expressions. Although the bootstrap test is designed for a data generating process with fixed parameters (i.e., independent of the sample size n), we also discuss how to obtain valid inference for sequences of DGPs with parameters approaching the boundary at the n−1/2 rate. A simulation study demonstrates that the new tests: (i) have excellent finite sample behaviour in terms of empirical rejection probabilities under the null as well as under the alternative; (ii) provide a useful complement to existing procedures based on Ljung-Box type approaches. Two data examples illustrate the implementation of the proposed tests in applications.

18:15 - 19:15 Poster Session
Room: Foyer

- **Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets**
Arnak Dalalyan, Lionel Riou-Durand, *Avetik Karagulyan*

  Abstract: In this paper, we provide non-asymptotic upper bounds on the error of sampling from a target density using three schemes of discretized Langevin diffusions. The first scheme is the Langevin Monte Carlo (LMC) algorithm, the Euler discretization of the Langevin diffusion. The second and the third schemes are, respectively, the kinetic Langevin Monte Carlo (KLMC) for differentiable potentials and the kinetic Langevin Monte Carlo for twice-differentiable potentials (KLMC2). The main focus is on the target densities that are smooth and log-concave on $\RR^p$, but not necessarily strongly log-concave. Bounds on the computational complexity are obtained under two types of smoothness assumption: the potential has a Lipschitz-continuous gradient and the potential has a Lipschitz-continuous Hessian matrix. The error of sampling is measured by Wasserstein-q distances and the bounded-Lipschitz distance. We advocate for the use of a new dimension-adapted scaling in the definition of the computational complexity, when Wasserstein-q distances are considered. The obtained results show that the number of iterations to achieve a scaled-error smaller than a prescribed value depends only polynomially in the dimension.

- **Bayesian Nonparametric Methods for Comparative Judgement Models and Vulnerability Estimation in Developing Countries**
Rowland Seymour, David Sirl, Simon Preston, James Goulding, Madeleine Ellis

  Abstract: Indicators for many UN Sustainable Development Goals, such as the prevalence of female genital mutilation and perinatal mortality rates, are often unreliable or unavailable in less developed countries. Comparative judgement models offer a solution, allowing respondents to leverage local knowledge, however, data is often difficult and expensive to collect. We have developed a Bayesian Nonparametric method to identify where the most vulnerable citizens in cities in less developed countries live. By incorporating a spatial structure into the model, we reduce the amount of data required. We develop a finite element Gaussian Process mixture model to identify where vulnerable areas in the city are located. We model the vulnerability of each area in the city through three spatial functions: the first model vulnerability on a city wide level, then second on a suburb level and the final on a neighbourhood level. The functions may require complex parametric forms to describe the variation in vulnerability, and given the data any choice we make will be challenging to justify. We instead use place a Gaussian Process prior distribution and learn these functions directly from the data. Current comparative judgement methods do not take a spatial element into account, and by doing so we can considerably reduce the amount of data required to estimate the vulnerability of each area. We demonstrate our method on a data set collected in Dar es Salaam, Tanzania, where we are able to identify a clear north-south divide in the city as well as highlight several slums in the city centre.

- **Mean Shrinkage Estimation for Diagonal Multivariate Natural Exponential Families**
Nikolas Siapoutis, Bharath Sriperumbudur, Donald Richards

  Abstract: Shrinkage estimators have been studied widely in statistics and have profound impact in many applications. Recently, Xie, et al. (2012, 2016) developed shrinkage estimators for the simultaneous estimation of the mean parameters of natural exponential families with quadratic variance functions (NEF-QVF). In this paper, we extend those results to the simultaneous estimation of the mean parameters of diagonal multivariate natural exponential families. In studying the family of distributions with quadratic diagonal covariance matrices, we propose a class of shrinkage estimators for the mean parameters and construct an unbiased estimator of the risk from estimating those parameters. Further, we establish the asymptotic optimality of the shrinkage estimators under squared error loss as n, the sample size tends to infinity. Under the assumption that both p, the dimension, and n tend to infinity in such a way that p/n goes

  to c for c in [0, 1), we again establish the asymptotic optimality of the shrinkage estimators. Finally, we consider the case of the diagonal multivariate natural exponential families, which include the multivariate normal, Poisson, gamma, multinomial and negative multinomial distributions, and we establish the asymptotic optimality under weaker assumptions.

- **A Gaussian model for survival data subject to dependent censoring and confounding**
Gilles Crommen, Jad Beyhum, Ingrid Van Keilegom

  Abstract: This paper considers the problem of inferring the causal effect of a variable Z on a survival time T. The error term of the model and Z are correlated, which leads to a confounding issue. Additionally, T is subject to dependent censoring, that is, T is right censored by a censoring time (C) which is dependent on T. In order to tackle the confounding issue, we leverage a control function approach relying on an instrumental variable W. It is assumed that T and C follow a joint regression model with bivariate Gaussian error terms and an unspecified covariance matrix, which allows us to handle dependent censoring in a flexible manner. We derive conditions under which the model is identifiable. A two-step estimation procedure is proposed and we show that the resulting estimator is consistent and asymptotically normal. Simulations are used to confirm the validity and finite-sample performance of the estimation procedure. Finally, the proposed method is applied to a bone marrow transplant data set. Non-parametric extensions of the model are being considered as well.

- **Advances in Multi-Output Quantile Regression**
  Miroslav Siman

  Abstract: Single-response quantile regression has proved useful for modelling univariate conditional distributions, which makes its generalization to vector responses highly desirable. There has been quite a few different approaches to the problem due to the lack of canonical ordering in multivariate spaces. Three multi-output quantile regression methods and their recent developments are presented: ameboid quantile regression, directional multi-output quantile regression (and its recent application to testing axial symmetry), and elliptical quantile regression (and its recent single-response version called bi-quantile regression). Further details and references can be found in the following resources:

- **Partially linear additive models on symmetric positive-definite matrices and Lie groups**
  Changwon Choi, Byeong Uk Park, Zhenhua Lin

  Abstract: We propose an extension of the partially linear additive model on the Lie groups, such as the space of symmetric and positive-definite matrices with Log-Euclidean or Log-Cholesky metric, and develop a semiparametric regression methods on these spaces. These spaces have an abelian group structure that provides a connection with partially linear additive models in a tangent space. The method estimates the parametric and nonparametric components using profiling techniques. We show that, under appropriate assumptions, the estimators of parametric components have convergence rates of parametric models. We also prove that the estimators of nonparametric components have convergence rates and asymptotic normality of univariate nonparametric models despite the multi-dimensional covariates. We present several simulation studies to evaluate the numerical performance of the proposed method. The methods are illustrated with the diffusion tensor brain imaging data obtained from the Alzheimer's Disease Neuroimaging Initiative database and multiple covariates.

- **Reconstruction of discretely sampled functional data with missing values**
  Siegfried Hörmann, *Maximilian Ofner*

  Abstract: In this work, we consider the problem of reconstructing partially observed functional data, when functions are discretely sampled with additive noise. Exploiting an underlying factor structure, we propose a two-step algorithm for the imputation of the non-observed part. First, loadings are estimated by a PCA estimator which is based on a completely observed subset of the sample. Secondly, factor estimates are obtained by projecting the partially observed functions onto the loadings. It is shown that the corresponding estimator for the common component allows for consistent reconstruction of the true signal under relatively mild assumptions. The established convergence rates are uniform over both time and cross-section and hold without restrictive smoothness assumptions.

FRIDAY 24 JUNE 2022

9:00 - 10:00 Special Invited Talk: Gabor Lugosi

Chair: Omiros Papaspiliopoulos
Room: Akamas A

Problems in network archaeology: root finding and broadcasting

Abstract: Networks are often naturally modeled by random processes in which nodes of the network are added one-by-one, according to some random rule. Uniform and preferential attachment trees are among the simplest examples of such dynamically growing networks. The statistical problems we address in this talk regard discovering the past of the network when a present-day snapshot is observed. We present a few results that show that, even in gigantic networks, a lot of information is preserved from the very early days. In particular, we discuss the problem of finding the root and the broadcasting problem.

10:00 - 11:00 Contributed Paper Session 3

Nonparametric regression

Chair: Leonie Selk
Room: Akamas C

10:00 Nonparametric tests to detect trends based on theory of records. An application to the analysis of climate change
Ana C Cebrian, Jorge Castillo, Jesús Asín

Abstract: This work presents some nonparametric tests to assess the hypothesis of i.i.d. sequences of variables based on the theory of records. The suggested tests are based on three types of variables related to the occurrence of records, binary variables indicating the occurrence of a record at time t, the number of records up to time t and the times of occurrence of records in a period. All the tests are developed using an important property: under the classical record model, that is i.i.d. sequences of variables (Xt), monotone transformations of the variables do not affect the distribution of the variables related to the occurrence of records. This property ensures that the distribution of record times does not depend on the distribution of the original sequence (Xt). This property is very useful to develop nonparametric test related to the occurrence of records, both in order to obtain asymptotic distributions of the statistics and also to implement Monte Carlo and bootstrap procedures. A study of the size and power of the tests is performed in order to compare the properties of all of them. These tests are applied to assess the existence of trends (global warming) using series of records from daily temperature series in Spanish locations. The application of the tests in this type of data presents some problems, such as the existence of seasonal behavior, correlation and the scarcity of records. Some useful preprocessing tools are suggested to deal with these problems.

10:20 Panel nonparametric regression with tensorial long short-term memory recurrent neural networks
Andrej Srakar

Abstract: Neural networks and deep learning are current state-of-the-art in machine learning and artificial intelligence. Seldom have they been applied to panel data estimation, in particular as replacement for nonparametric fixed or random effects specification. This paper builds on a previous contribution by Crane-Droesch (2017) upgrading his approach using tensorial long short-term memory recurrent neural networks based on tensor Tucker decomposition. Neural networks are fitted to panel nonparametric fixed effects specification. Loss function and backpropagation are defined in accordance with Kolda and Bader (2009). The model is estimated by a variant of minibatch gradient descent using approximate inference via linear Taylor expansion. Model performance is compared to tensorial gated recurrent units and parametric and nonparametric fixed and random effects models. To evaluate the novel estimator we use asymptotic approaches for neural network estimation and Monte Carlo type simulation evidence. Extension of the approach to nonparametric random effects specification is discussed. The approach is applied to the prediction of agricultural yields from weather data, relevant for short-term economic forecasts as well as for longer-range climate change impact assessment. The model is general and could be used for high-dimensional regression adjustment, general nonparametric regression problems, heterogeneous treatment effect estimation, and forecasting with longitudinal data.

10:40 Nonparametric Regression and Classification with Functional, Categorical, and Mixed Covariates
Leonie Selk, Jan Gertheiss

Abstract: We consider nonparametric prediction with multiple covariates, in particular categorical or functional predictors, or a mixture of both. The method proposed bases on an extension of the Nadaraya-Watson estimator where a kernel function is applied on a linear combination of distance measures each calculated on single covariates, with weights being estimated from the training data. The dependent variable can be categorical (binary or multi-class) or continuous, thus we consider both classification and regression problems. The methodology presented is illustrated and evaluated on artificial and real world data. Particularly it is observed that the data-driven weights downgrade those covariates that are irrelevant, noise variables whereas relevant covariates are weighted distinctly higher. Thus variable selection is automatically performed and prediction accuracy is increased.

Survival analysis II

Chair: Ali Shariati
Room: Aphrodite A

10:00 Instrumental variable quantile regression under random right censoring
Lorenzo Tedesco, Jad Beyhum, Ingrid Van Keilegom

Abstract: The paper presents a new estimator for quantile regression analysis in the presence of endogenous variables and censored data, using instrumental variables. The structural quantile is supposed to be linear in the covariate, the censoring variable to be independent of the outcome of interest, and no assumptions on the covariate distribution are required. Results of consistency and asymptotic distribution are provided, together with simulations and empirical application of the proposed method.

10:20 Single-index mixture cure models. An application to a study of cardiotoxicity in breast cancer patients.
Beatriz Piñeiro Lamas, Ricardo Cao, Ana López Cheda

Abstract: Standard survival models assume that, in the absence of censoring, all individuals will experience the event of interest. However, sometimes this is not realistic. For example, if we consider cancer patients being treated and the event is the appearance of an adverse effect, there will be patients that will never experience it. Those who will never develop this health condition will be considered as cured. To incorporate this cure fraction, classical survival analysis has been extended to cure models. In particular, mixture cure models allow to estimate the probability of being cured and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates. We fill this important gap by considering both vector and functional covariates and proposing a single-index model for dimension reduction. The methodology is applied to a cardiotoxicity dataset from the University Hospital of A Coruña.

10:40 A Goodness-of-fit Test with Length-biased Data
Ali Shariati, Vahid Fakoor, Mahboobeh Akbari

Abstract: Length-biased data arises in many disciplines such as survival analysis, econometric, renewal processes, biomedicine, and physics. In this article, a research into the bootstrap approach using the Cramér–von Mises statistic is conducted to obtain a goodness–of–fit test which is based on a sample from a length–biased distribution. The proposed method is an extension of the leveraged bootstrap approach for data collected through a length-biased sampling procedure. The limiting distribution of the proposed test statistic is derived and revealed to be asymptotically distribution free. The power of the proposed one-sample test is evaluated and compared with an existing alternative test through a simulation study. The proposed method is applied for a set of real data on automobile brake pads.

Statistical inference in complex models

Chair: Yann Issartel
Room: Aphrodite B

10:00 Geometric-Median-of-Means in Non-Positive Curvature Spaces
Ho Yun, Byeong Uk Park

Abstract: In Euclidean spaces, the empirical mean vector as an estimator of the population mean is known to have polynomial concentration unless a strong tail assumption is imposed on the underlying probability measure. The idea of median-of-means tournament has been considered as a way of overcoming the sub-optimality of the empirical mean vector. In this talk, to address the sub-optimal performance of the empirical mean in a more general setting, we consider general Polish spaces with a general metric, which are allowed to be non-compact and of infinite-dimension. We discuss the estimation of the associated population Fréchet mean, and for this we extend the existing notion of median-of-means to this general setting. We devise several new notions and inequalities associated with the geometry of the underlying metric, and using them we study the concentration properties of the extended notions of median-of-means as the estimators of the population Fréchet mean. We show that the new estimators achieve exponential concentration under only second moment condition on the underlying distribution, while the empirical Fréchet mean has polynomial concentration. We focus our study on spaces with non-positive Alexandrov curvature since they afford slower rates of convergence than spaces with positive curvature. We note that this is the first work that derives non-asymptotic concentration inequalities for extended notions of the median-of- means in non-Euclidean spaces with a general metric.

10:20 Bootstrap inference in functional linear regression models with scalar response
Hyemin Yeon, Danieal Nordman, Xiongtao Dai

Abstract: In fitting linear regression models for functional data, a complicating factor with regressors as random curves is that regression estimators have complex distributions, due to issues in bias and scaling. Bias arises because the target slope function is infinite-dimensional, while finite-sample estimators necessarily involve truncations. To approximate sampling distributions, we develop a residual bootstrap method. Despite the parametric regression problem, the bootstrap for functional data requires a development that resembles resampling for nonparametric regression with multivariate regressors. Essentially, original- and bootstrap-data estimators require coordination in the truncation levels to remove bias (akin to tuning parameter choices). The resulting bootstrap has wide applicability for

constructing both confidence and prediction regions at target regressor points, and with coverage properties even holding conditionally on data regressors; the method also extends to simultaneous regions. Establishment of the bootstrap further involves correcting and generalizing a foundational central limit theorem for functional linear regression. Numerical studies verify our theory, showing that the bootstrap performs better than normal approximations, and also suggest a rule of thumb for setting the truncation levels. The bootstrap method is illustrated with an application to wheat spectrum data.

10:40 The Seriation and 1D-localization problems in latent space models.
Yann Issartel

Abstract: Motivated by applications in archeology for relative dating of objects, or in 2D-tomography for angular synchronization, we consider the problem of statistical seriation where one seeks to reorder a noisy disordered matrix of pairwise affinities. This problem can be recast in the powerful latent space terminology where the affinity between a pair of items is modeled as a noisy observation of a function f(x_i,x_j) of the latent points x_i, x_j of the two items in a one-dimensional space. This reformulation naturally leads to the problem of estimating the latent positions in the latent space. Under non-parametric assumptions on the affinity function f, we introduce a procedure that provably localizes all the latent positions with a maximum error of the order of the square root of log(n)/n. This rate is proven to be minimax optimal. Different computationally efficient procedures are also analyzed, under different set of assumptions. Our general results can be instantiated to the original problem of statistical seriation, leading to new bounds for the maximum error in the ordering.

Directional data

Chair: Diego Bolon

Room: Christian Barnard

10:00 Nonparametric regression estimation for a functional-circular model
Andrea Meilan-Vila, Rosa M. Crujeiras, Mario Francisco-Fernández

Abstract: The analysis of a variable of interest which depends on other variable(s) is a typical issue appearing in many practical problems. Regression analysis provides the statistical tools to address this type of problems. This topic has been deeply studied, especially when the variables in study are of Euclidean type. However, there are situations where the data present certain kind of complexities, for example, the involved variables are of circular or functional type, and the classical regression procedures designed for Euclidean data may not be appropriate. In these scenarios, these techniques would have to be conveniently modified to provide useful results. This work aims to design and study a new approach to deal with regression function estimation for models with a circular response and a functional covariate. The asymptotic bias and variance of the proposed estimator are calculated. Some guidelines for its practical implementation are provided, checking its sample performance through simulations. Finally, the behavior of the estimator is also illustrated with a real data set.

10:20 Nonparametric multimodal regression for a circular response
Maria Alonso-Pena, Rosa M. Crujeiras

Abstract: Many experiments concerning animal behavior focus on the study of escape strategies subject to different predictor variables. The datasets resulting from such investigations can usually be analyzed from a regression perspective, but taking into account that the response variable, escape direction, is a circular variable. Circular data are defined on the unit circumference, and their peculiar nature makes usual statistical techniques not suitable for a proper analysis. This has lead to a considerable increase of attention on methods specifically tailored for this kind of data in recent years. Although parametric and nonparametric regression models for a circular response have been proposed in the statistical literature, these models regard the regression function as the circular mean direction conditioned to the value of the covariates. However, estimating the conditional mean is not always appropriate when dealing with animal escape data, which usually presents a multimodal structure. In this work, a new nonparametric regression method to deal with this kind of data is presented, in which the conditional local modes are estimated instead of the conditional mean direction. The new approach is based on the maximization of the conditional kernel density estimator for directional variables. The maximization is carried out by employing the circular mean shift algorithm, which can be regarded as a gradient ascent-type algorithm in the unit circumference. Asymptotic error rates for the new estimator are derived, and its performance in practice is assessed through a simulation study. A real data example concerning the escape behavior of larval zebrafish is used to illustrate the proposed methodology.

10:40 A hybrid method for estimating highest density regions of directional data
Diego Bolón, Rosa M. Crujeiras, Alberto Rodríguez-Casal

Abstract: Highest density regions (HDRs) are defined as sets where the density function takes relatively large values. Given a data sample, the estimation of HDRs of the underlying density is required for data modelling, exploration and visualization. For example, it was found to be useful tool for approximating the localization of minefields based on aerial observations, analyzing seismic data and detecting outliers within a sample. Estimating HDRs for Euclidean data (uni or multidimensional) has been widely considered in the statistical literature. However, HDRs estimation for directional data (i.e. data on the sphere or hypersphere) has not been approached until very recently. The most natural route for estimating directional HDRs is a plug-in method: the HDR estimator is defined directly as the HDR of an estimator of the underlying density. Although this estimation method is conceptually simple, it does not consider any information about the shape of the HDRs. This is not problematic when we do not have any prior geometric information about the theoretical HDRs, but it may lead to inconsistent estimation of HDRs otherwise. If we know that the theoretical HDRs are convex or connected, it makes sense to include that information in the estimation method. Keeping this in mind, we propose a new non-parametric HDR estimator for directional data that takes into account the geometry of the HDRs of the underlying density. Specifically, our estimation technique is a hybrid method that combines information of the kernel density estimator with smoothness assumptions on the considered class of sets.

Robust Statistics II

Chair: Markus Neuhaeuser

Room: Leda

10:00 **Computation of the halfspace depth regions – exact algorithm**
Petra Laketa, Stanislav Nagy, Vit Fojtik, Pavlo Mozharovskyi

Abstract: The halfspace depth central regions are an analogue of quantiles for multivariate data. We are concerned with their exact computation, and analyse the recent algorithm implemented in the R package TukeyRegion. Even though that algorithm is not exact in general, we proved its exactness in some special cases and used those results to construct an exact algorithm. Our algorithm computes simultaneously all the depth regions starting from the one with the lowest level up to a given constant level.

10:20 **Usefulness of the square-root allocation rule for many-to-one comparisons under non-normality**
Markus Neuhaeuser

Abstract: In a many-to-one situation, i.e. when comparing a control group with k other groups, Dunnett's test is recommended for normally distributed data. In this situation unbalanced sample sizes, chosen according to the square-root sampling allocation rule, can give a more powerful test than a balanced design, see e.g. Neuhäuser et al. (2021). Herberich et al. (2010) proposed a robust procedure for many-to-one comparisons. For this method, no assumptions regarding distribution or variance homogeneity are necessary. Here, we investigate whether the square-root allocation rule is also useful in case of non-normality with or without variance homogeneity between groups. Herberich, E., Sikorski, J. & Hothorn, T. (2010): A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. PLoS One 5(3), e9788. Neuhäuser, M., Mackowiak, M. & Rixton, G.D. (2021): Unequal sample sizes according to the square-root allocation rule are useful when comparing several treatments with a control. Ethology 127(12), 1094-1100.

11:00 - 11:30 Coffee Break

11:30 - 12:30 Contributed Paper Session 4

Permutation tests

Chair: Nick Koning

Room: Akamas C

11:30 **Time series comparisons: a permutation approach**
Stefano Bonnini, Michela Borghesi

Abstract: In recent years, the literature on univariate and multivariate statistical methods for the analysis of groups of time series has grown. The contributions focus mainly on problems of clustering. The literature seems instead lacking in contributions concerning comparisons of groups of time series for inferential purposes. This work intends to fill the gap with regard to problems of hypothesis testing. In particular, we address the problem of comparing two or more groups of time series in a given time period, in order to test the hypothesis of systematic differences (not due to sample variability) in the trajectories over time, at least during one or more sub-periods of the considered time span. For instance, we may be interested in comparing the behavior over time of the returns of two different types of listed companies (e.g. sustainable and non-sustainable enterprises), to test the hypothesis of equality of the financial performance of the two types of companies over time. The main difficulty is related to the usually high length of the considered time span that makes the test classifiable as a Big Data problem. The proposed solution is based on the combination of permutation tests and satisfies, among others, an important property in this framework. In fact, the power of the combined test is a non-decreasing function of the number of considered time points.
This work is supported by the University of Ferrara which funded the project "Measuring and evaluating social sustainability, inclusion, and accessibility in a University: methods and applications" (Fund for the Promotion of Research, FIR-2020).

11:50 **More Efficient Exact Permutation Tests**
Nick Koning, Jesse Hemerik

Abstract: Non-parametric tests based on permutation, rotation or sign-flipping are examples of so-called group-invariance tests. These tests rely on invariance of the null distribution under a set of transformations that has a group structure, in the algebraic sense. Such groups are often huge, which makes it computationally infeasible to use the entire group. Hence, it is standard practice to test using a randomly sampled set of transformations from the group. This random sample still needs to be substantial to obtain good power and replicability. We improve upon the standard practice by using a well-designed subgroup of transformations instead of a random sample. We show this can yield a more powerful and fully replicable test with the same number of transformations. For a normal location model and a particular design of the subgroup, we show that the power improvement is equivalent to the power difference between a Monte Carlo Z-test and Monte Carlo t-test. In our simulations, we find that our test has the same power as a test based on sampling that uses double the number of random transformations. These benefits come entirely `for free', as our methodology relies on an assumption of invariance under the subgroup, which is implied by invariance under the entire group.

Change point analysis

Chair: Florian Pein
Room: Aphrodite A

11:30 Semiparametric Detection of Changepoints in Location, Scale and Copula
Gaurav Agarwal, Idris Eckley, Paul Fearnhead

Abstract: This research proposes a new method to detect changepoints in the location and scale of univariate data sequences. The proposed method assumes that the data belongs to the location-scale family of distributions and estimates the associated densities non-parametrically. Specifically, the approach does not require knowledge of the functional form of the distribution of the data sequence. As such, the approach can detect changepoints in many distributions. We also propose a new method to detect changes in the location of multivariate sequences, using the marginals and a copula to capture the dependence between variables without the influence of marginal distributions. The performance of the proposed semiparametric approach is contrasted against both other competing nonparametric and Gaussian methods, via simulation studies, as well as applications arising from health and finance.

11:50 NP-FOCuS: a Nonparametric Approach for Online Changepoint Detection
Gaetano Romano, Idris Eckley, Paul Fearnhead

Abstract: Online changepoint detection aims to detect anomalies and changes in real-time in high-frequency data streams, sometimes with limited available computational resources. This is an important task that is rooted in many real-world applications, including and not limited to cybersecurity, medicine and astrophysics. While fast and efficient online algorithms have been recently introduced, these rely on Gaussianity assumptions which are often violated in practical applications. To allow for a wider range of scenarios, we consider a nonparametric approach to detect any change in the distribution of a data stream. Our procedure, NP-FOCuS, has a computational cost that is log-linear in the number of observations and is suitable for high-frequency data streams. In terms of detection power, NP-FOCuS is seen to outperform current nonparametric changepoint techniques in different settings. We demonstrate the utility of the procedure in monitoring both simulated and industrial data.

12:10 Cross-validation for change-point regression: pitfalls and solutions
Florian Pein

Abstracts: Cross-validation is the standard approach for tuning parameter selection in many non-parametric regression problems. However its use is less common in change-point regression, perhaps as its prediction error-based criterion may appear to permit small spurious changes and hence be less well-suited to estimation of the number and location of change-points. We show that in fact the problems of cross-validation with squared error loss are more severe and can lead to systematic under- or over-estimation of the number of change-points, and highly suboptimal estimation of the mean function in simple settings where changes are easily detectable. We propose two simple approaches to remedy these issues, the first involving the use of absolute error rather than squared error loss, and the second involving modifying the holdout sets used. For the latter, we provide conditions that permit consistent estimation of the number of change-points for a general change-point estimation procedure. We show these conditions are satisfied for optimal partitioning using new results on its performance when supplied with the incorrect number of change-points. Numerical experiments show that the absolute error approach in particular is competitive with common change-point methods using classical tuning parameter choices when error distributions are well-specified, but can substantially outperform these in misspecified models.

Robust statistics III

Chair: Ragnhild Laursen
Room: Aphrodite B

11:30 Robust estimation of a regression function in exponential families
Juntong Chen

Abstract: We consider the problem of estimating a regression function when the distribution of the data is modelled by an exponential family. Several interesting problems are under this setting, for example logit, Poisson and exponential regressions. Our estimation strategy is based on Rho-estimation and we present a non-asymptotic exponential inequality for the deviation of their risk. We deduce from this inequality that the estimator is robust to contamination, the presence of outliers and model misspecification. We also provide a uniform risk bound over the class of Hölderian functions and prove the optimality of the estimator over this class up to a logarithmic factor. Finally, we carry out a simulation study in order to compare the performance of Rho-estimators to the maximum likelihood estimator and median-based ones.

11:50 Robust estimation of non-negative matrix factorization for mutational signatures using a flexible parametrization
Ragnhild Laursen, Asger Hobolth, Lasse Maretty Sørensen

Abstract: Mutational signatures are derived from somatic mutations in the cancer genome using non-negative matrix factorization (NMF). The method factorizes the data of mutational counts into two non-negative matrices containing the mutational signatures and the corresponding weights, where each entry in the two matrices is a free parameter to be estimated. The resulting factorization is clearly not unique up to permutation and scaling, but there could also exist other less obvious transformations that make the solution non-unique.

To avoid this non-uniqueness problem and reduce the chance of overfitting we are parametrizing the matrix containing the mutational signatures. Mutational signatures depends on both the base mutation and flanking nucleotides, which can be seen as different features. They have been parametrized using only first-order interactions, but we introduce a more flexible and novel framework that allows inclusion of only the important interactions between the base mutation and the flanking nucleotides. In particular we look at the second-order interactions. We argue that the second-order interaction signatures are biologically plausible, and demonstrate that they are statistically stable and adequate. Second-order interaction signatures often strike the right balance between appropriately fitting the data and avoiding overfitting. They provide a better fit to data and are biologically more realistic than first-order signatures, and the parametrization is more stable than the parameter-rich three-way interaction signatures.

Applications of nonparametric inference

Chair: Prajamitra Bhuyan
Room: Christian Barnard

11:30 Adaptive wavelet estimation of a latent variable model
Andrej Srakar, Marilena Vecco

Abstract: Latent variable models provide statistical tool for explaining and analyzing underlying structure of multivariate data by using the idea that observable phenomena are influenced by underlying factors which cannot be observed or measured directly. One possibility to fit them is to assume that the underlying distribution is Gaussian, and therefore it is uniquely determined by its covariance structure. This is commonly done using maximum likelihood and works under large sample asymptotics. In a recent article, first in developing non-parametric regression with latent variables, Kelava et al. (2017) used a two-step approach to fit a non-parametric regression model: in the first step they have fitted a common factor analysis model and then applied B-spline non-parametric regression techniques to analyze the relation between the latent variables. Following this approach, we extend their article in multiple directions. Common factor analysis as key part of the approach is fit using nonparametric Bayesian approach following Piatek and Papaspiliopoulos (2018) and Knowles and Ghahramani (2011), allowing for correlated factors. Moreover, we extend the spline approach of Kelava et al. to adaptive (block-thresholded) wavelets which were shown to have good finite sample properties (Cai, 1999; 2009). This allows the estimation to be used for smaller samples. We derive asymptotic properties of the approach and show the estimator is consistent, asymptotically normal and efficient. The behaviour of the estimator is studied in Monte Carlo simulation comparing it to the generally used Latent Moderated Structural Equations (LMS) and Structural Equation Mixture Modeling (SEMM) estimators as well as to the Kelava et al. spline estimator. Short application studies relationship of life satisfaction and financial indicators of older people using data of Survey of Health, Ageing and Retirement in Europe (SHARE).

11:50 Sales Forecasting with Weather Data using Fuzzy Natural Logic
Tomas Tichy

Abstract: Reliable estimation of customer demand for products and services constitutes a key aspect of financial planning in every company. When estimating future sales, as a proxy to demand, in addition to pure economic quantities, a large selection of (exogenous) variables specific to a given product can be considered. Potential impact of weather conditions on sales has been known for very long time, though the research using weather data has been mostly focused on energy sector. As concerns retail, several authors have started to analyze this issue only recently. This paper proposes a novel approach studying (quarterly) sales using weather data and is inspired by fuzzy natural logic. The method is based on modeling the influence of average temperatures on sales by fuzzy linguistic IF-THEN rules. The proposed methodology is applied to real data of quarterly ice-cream sales and compared with a standard approach. The results are promising especially when monthly average temperatures are considered.

12:10 Analysing the causal effect of London cycle superhighways on traffic congestion
Prajamitra Bhuyan, Emma McCoy, Haojie Li, Daniel Graham

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

Extreme values

Chair: Dora Prata Gomes
Room: Leda

11:30 Semi-parametric weighted Hill estimators

Frederico Caeiro, *Ayana Mateus*

Abstract: In statistics of extremes, the extreme value index dominates the tail behaviour and needs to be estimated in a precise way because other tail parameters such as an extreme quantile or a tail probability depends on such value. In this work our focus is on the estimation of a strictly positive extreme value index from a model with a Pareto-type right tail. Under this semi-parametric framework we propose a new class of weighted Hill estimators, parameterized with a tuning parameter. We analyse the asymptotic limiting distribution of this new class of estimators assuming the validity of a second order framework and illustrate their performance with a Monte Carlo simulation study. A comparison with other important estimators from the literature are also provided.

11:50 Nonparametric Asymptotic Confidence Intervals for Extreme Quantiles
Samuel Maistre, Laurent Gardes

Abstract: We propose new asymptotic confidence intervals for extreme quantiles. Our intervals, to some extent, generalize the idea used by Weissmann (1978) in the context of the point estimation of extreme quantiles. We also propose a bias-reduction procedure to improve the coverage probability of our intervals. Using simulations, our intervals' performances are investigated and compared to those proposed by Buitendag, Beirlant, and de Wet (2020).

12:10 Nonparametric resampling methods in the estimation of parameters of rare events
Dora Prata Gomes, Helena Penalva, Sandra Nunes, Manuela Neves

Abstract: Extreme Value Theory has been asserting itself as one of the most important statistical
theories for the applied sciences providing a solid theoretical basis for deriving statistical models describing extreme or even rare events. The efficiency of the inference and estimation procedures depends on the tail shape of the distribution underlying the data. Computer-intensive methods, which emerged when computers became more powerful have been developed in the last decades. The most well known are perhaps the jackknife (\cite{Quenouille1949}) and the bootstrap (\cite{Efron1979}) methodologies. These two methodologies have been used with success in Extreme Value Theory overcoming the difficulties that appear in the semi-parametric estimation of parameters of extreme events. The main focus of this work is to perform an univariate extreme value analysis illustrating and applying computational nonparametric methods, such as the Generalized Jackknife and the Bootstrap that have revealed to improve the parameter estimators. Different approaches for resampling need to be considered depending on whether we are in an independent or in a dependent setup. Some recent estimators of the tail index are also compared and relevant parameters such as {\it high quantile}, the {expected shortfall}, the {\it return period} of a high level are also studied. A practical application on the effect of taking into consideration or not the choice of the tail of the underlying distribution and consequently the adequate extreme value index estimation is presented.

12:30 - 13:30 Lunch Break

13:30 - 15:30 Invited Paper Session 9

Inference for Dependent Data

Organiser: Kostas Fokianos

Chair: Kostas Fokianos

Room: Akamas A

13:30 High-Dimensional Mixed Models with Varying Coefficients and Functional Random Effects
Michael Law, *Ya'acov Ritov*

Abstract: We consider a sparse high-dimensional varying coefficients model with random effects, a flexible linear model allowing covariates and coefficients to have a functional dependence with time. For each individual, we observe discretely sampled responses and covariates as a function of time as well as time-invariant covariates. Under sampling times that are either fixed and common or random and independent amongst individuals, we propose a projection procedure for the empirical estimation of all varying coefficients. We extend this estimator to construct confidence bands for a fixed number of varying coefficients.

14:00 Monotonic Alpha-divergence Minimisation for Variational Inference
Kamelia Daudel, *Randal Douc*, Francois Roueff

Abstract: In this paper, we introduce a novel family of iterative algorithms which carry out alpha-divergence minimisation in a Variational Inference context. They do so by ensuring a systematic decrease at each step in the alpha-divergence between the variational and the posterior distributions. In its most general form, the variational distribution is a mixture model and our framework allows us to simultaneously optimise the weights and components parameters of this mixture model. Notably, our approach permits to build on various methods previously proposed for alpha-divergence minimisation such as Gradient or Power Descent schemes and we also shed a new light on an integrated Expectation Maximization algorithm. Lastly, we provide empirical evidence that our methodology yields improved results on several multimodal target distributions.

14:30 Hold-out estimates of prediction models for Markov processes
Joseph Rynkiewicz

Abstract: We consider the selection of prediction models for Markovian time series. For this purpose, we study the theoretical properties of the hold-out method. In the econometrics literature, the hold-out method is called ``out-of-sample'' and is the main method to select a suitable time series model. This method consists of estimating models on a learning set and picking up the model with minimal empirical error on a validation set of future observations. Hold-out estimates are well studied in the independent case, but, as far as we know, this is not the case when the validation set is not independent of the learning set. In this paper, assuming uniform ergodicity of the Markov chain, we state generalization bounds and oracle inequalities for such method; in particular, we show that the ``out-of-sample'' selection method is adaptative to noise condition.

15:00 **Testing for changes in the tail-index of Long Memory Stochastic Volatility time series**
Davide Giraudo, *Annika Betken*, Rafal Kulik

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

Nonparametric Approaches to Complex Data Problems

Organiser: Somnath Datta

Chair: Somnath Datta

Room: Akamas C

13:30 **A Bayesian nonparametric approach to causal mediation with multiple mediators**
Michael Daniels

Abstract: We introduce an approach for causal mediation with multiple mediators. We model the observed data distribution using a new Bayesian nonparametric approach that allows for flexible default specifications for the distribution of the outcome and the mediators conditional on mediator/outcome confounders. We briefly explore the properties of this specification and introduce assumptions that allow for the identification of direct and both joint and individual indirect effects. We use this approach to examine the effect of antibiotics as mediators of the relationship between bacterial community dominance and ventilator associated pneumonia and conduct simulation studies to better understand the frequentist properties of our approach.

14:00 **Structurally Sparse Bayesian Neural Networks**
Taps Maiti, Sanket Jantree, Shrijita Bhttacharya

Abstract: Network complexity and computational efficiency are increasingly significant aspects of deep learning. Sparse deep learning addresses these challenges by recovering the sparse structure of target functions while reducing over-parameterized model to a compact size. In this work, we adopt Bayesian solution through spike-and-slab group shrinkage priors to structurally reduce the network by pruning excess nodes. We propose variational Bayes inferences with continuous relaxation of discrete variables for posterior approximation. It helps to circumvent the computational challenges of traditional Markov Chain Monte Carlo (MCMC) implementation. We establish variational posterior contraction rates along with the characterization of prior parameters under relaxed layer-wise number of nodes and coefficient bounds. With a layer-wise characterization of prior inclusion probabilities, we establish optimal contraction rates of the variational posterior for smooth functions. The numerical investigation demonstrates competitive predictive performance of our proposed models while achieving significant network compression thereby improving computational complexity during inference.

14:30 **Semiparametric Analysis of Clustered Interval-Censored Survival Data using Soft Bayesian Additive Regression Trees**
Debajyoti Sinha, Piyali Basak, Antonio Linero

Abstract: Popular parametric and semiparametric hazards regression models for clustered survival data are inappropriate and inadequate when the unknown effects of different covariates and clustering are complex. This calls for a flexible modeling framework to yield efficient survival prediction. Moreover, for some survival studies involving time to occurrence of some asymptomatic events, survival times are typically interval censored between consecutive clinical inspections. In this article, we propose a robust semiparametric model for clustered interval-censored survival data under a paradigm of Bayesian ensemble learning, called Soft Bayesian Additive Regression Trees or SBART (Linero and Yang, 2018), which combines multiple sparse (soft) decision trees to attain excellent predictive accuracy. We develop a novel semiparametric hazards regression model by modeling the hazard function as a product of a parametric baseline hazard function and a nonparametric component that uses SBART to incorporate clustering, unknown functional forms of the main effects, and interaction effects of various covariates. In addition to being applicable for left-censored, right-censored, and interval-censored survival data, our methodology is implemented using a data augmentation scheme which allows for existing Bayesian backfitting algorithms to be used. We illustrate the practical implementation and advantages of our method via simulation studies and an analysis of a prostate cancer surgery study where dependence on the experience and skill level of the physicians leads to clustering of survival times. We conclude by discussing our method's applicability in studies involving high dimensional data with complex underlying associations.

15:00 Regression Analysis of a Future State Entry Time Distribution Conditional on a Past State Occupation in a Progressive Multistate Model
Somnath Datta

Abstract: We present a nonparametric method for estimating the conditional future state entry probabilities and distributions of state entry time conditional on a past state visit when data are subject to dependent censorings in a progressive multistate model where Markovity of the system is not assumed. These estimators are constructed using the competing risk techniques with risk sets consisting of fractional observations and inverse probability of censoring weights. The fractional observations correspond to estimates of the numbers of persons who ultimately enter a state from which the future state in question can be reached in one step. We then address the corresponding regression problem by combining these marginal estimators with the pseudo-value approach. Performance of our regression scheme is studied using a comprehensive simulation study. An analysis of a well known existing data on graft-versus-host disease for bone marrow transplant individuals is presented using our novel methodology.

Computational Statistics

Organiser: Stefan Sperlich

Chair: Stefan Sperlich

Room: Aphrodite A

13:30 Nonparametric inference for big-but-biased data
Ricardo Cao, Laura Borrajo

Abstract: It is often believed that in a Big Data context, given the large amount of data available, the data reflect precisely the underlying population. However, the data are often strongly biased due to the procedure used for obtaining them. In order to reduce the signifficant bias that may appear in Big Data (Big-but-Biased Data, B3D), different testing methods for bias detection are used and completely nonparametric estimation methods for bias correction are proposed for large-sized but possibly biased samples. Nonparametric estimators for the mean of a transformation of the random variable of interest are considered. When ignoring the biasing weight function, two different setups are proposed. In Setup 1 (Borrajo and Cao (2021)) a small-sized simple random sample of the real population is assumed to be additionally observed, while in Setup 2 it is assumed that a twice biased sample of small size is observed. The asymptotic properties of the proposed estimators are extensively studied under suitable limit conditions on the small and the large sample sizes and standard and non-standard asymptotic conditions on the two bandwidths. The performance of the proposed nonparametric estimators is compared with the classical estimators based on the two samples involved in each setup through Monte Carlo simulation studies. Simulation results show that the new mean estimators outperform the classical empirical means for suitable choices of the two smoothing parameters involved. The influence of these smoothing parameters on the performance of the nal estimators is also studied, exhibiting a striking limit behaviour of their optimal values. In addition, bootstrap bandwidth selection methods for each nonparametric mean estimator are introduced. Finally, the proposed techniques are applied to several real data sets from different areas (see, for instance Borrajo and Cao (2020)).

14:00 Local inference for data giants
Gilles Cattani, *Michael Scholz*, Stefan Sperlich

Abstract: The advent of the age of big data brings new opportunities for modern societies and poses interesting challenges to statisticians. In this paper, we consider the case of data collected and stored in a decentralized manner which is, however, too large to be merged on a single platform and/or legally protected what excludes aggregation. For example, observations from multiple states, data generated by specific methodological stations, or information from different hospitals could be stored on different servers that we will call data giants. For data analysis, we adapt the local-polynomial regression technique with local bandwidths to this situation and allow further for a local variable selection of the LASSO-type. The proof of concept and computational details are given in a simulation study. An empirical application to data giants shows the practical relevance of the new algorithm.

14:30 Multivariate nonparametric change point detection with Random Forests and other classifiers
Solt Kovács, Malte Londschien, Peter Bühlmann

Abstract: We propose a novel view on nonparametric change point detection based on classifiers. We construct a log-likelihood type statistic that uses in-sample predictions for class probabilities. As specific examples we discuss Random Forests and k-nearest neighbours. While Random Forest classifiers lead to excellent empirical performance in multivariate (or possibly even high-dimensional) setups, they are computationally expensive to fit. Thus, we focus on approaches that rely only on a relatively small number of fits while maintaining the performance. For example, we utilise the advantages of the recently proposed Seeded Binary Segmentation method (Kovács et al., 2020) in our nonparametric scenario. We show extensive simulation results and also some supporting theoretical results for our nonparametric change point detection methodology.

15:00 Sparse Deep Learning
Johannes Lederer

Abstract: Sparsity is popular in statistics and machine learning, because it can avoid overfitting, speed up computations, and facilitate interpretations. In deep learning, however, the full potential of sparsity still needs to be explored. This presentation first recaps sparsity in the framework of high-dimensional statistics and then introduces sparsity-inducing methods and corresponding theory for modern deep-learning pipelines.

Multivariate Non-Parametric Tests

Organiser: Simos Meintanis

Chair: Charl Pretorius

Room: Aphrodite B

13:30 **On the IPCW approach for testing independence,** Marija Cuparić, *Bojana Milošević*

Abstract: Here we present a novel IPCW adaptation of the Kochar-Gupta (KG) test of independence, in the case of bivariate randomly censored data. Three different censoring schemes are considered: one of the targeted variables is censored, both targeted variables are censored with the same censoring variable and both targeted variables are censored with different censoring variables. The limiting properties of the test statistic are explored. In order to compare tests with a few well-known competitors, in terms of powers, several resampling procedures have been utilized to approximate null distribution. Special attention is given to comparison with a classical adaptation of the KG test related to the IPCW adaptation of U-statistics.

14:00 **Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility**
Tom Berrett

Abstract: Given a set of incomplete observations, we study the nonparametric problem of testing whether data are Missing Completely At Random (MCAR). Our first contribution is to characterise precisely the set of alternatives that can be distinguished from the MCAR null hypothesis. This reveals interesting and novel links to the theory of Frechet classes (in particular, compatible distributions) and linear programming, and we leverage tools developed in these fields to propose MCAR tests that are consistent against all detectable alternatives. Moreover, we define a natural measure of ease of
detectability (an incompatibility index), and exploit ideas from max-flow min-cut theory to prove that our tests achieve the optimal minimax separation rate according to this measure in certain cases.

14:30 **Inference procedures for models with multivariate elliptically symmetric stable Paretian laws**
Charl Pretorius, Simos Meintanis, John Nolan, Zhou Zhou

Abstact: We consider estimation and goodness-of-fit testing methods for multivariate symmetric location-dispersion stable Paretian random vectors in arbitrary dimension. The methods are based on the empirical characteristic function and are relatively easy to implement. Asymptotic properties of the proposed procedures are presented, while the favourable finite-sample properties are illustrated by means of a Monte Carlo study. The procedures are also applied to more complicated models involving such distributions (such as GARCH models) by using real data from the financial markets, both univariate and multivariate.

Functional estimation in Monte Carlo methods

Organiser: Sylvain Le Corff

Chair: Luc Lehericy

Room: Christian Barnard

13:30 **Diffusion Schrodinger Bridge and Score-Based Generative Modeling**
Valentin De Bortoli, James Thornton, Arnaud Doucet, Jeremy Heng

Abstract: Progressively applying Gaussian noise transforms complex data distributions to approximately Gaussian. Reversing this dynamic defines a generative model. When the forward noising process is given by a Stochastic Differential Equation (SDE), demonstrate how the time inhomogeneous drift of the associated reverse-time SDE may be estimated using score-matching. A limitation of this approach is that the forward-time SDE must be run for a sufficiently long time for the final distribution to be approximately Gaussian while ensuring that the corresponding time-discretization error is controlled. In contrast, solving the Schrodinger Bridge (SB) problem, i.e. an entropy-regularized optimal transport problem on path spaces, yields diffusions which generate samples from the data distribution in finite time. We present Diffusion SB (DSB), an original approximation of the Iterative Proportional Fitting (IPF) procedure to solve the SB problem, and provide theoretical analysis along with generative modeling experimentd. Beyond generative modeling, DSB offers a computational optimal transport tool as the continuous state-space analogue of the popular Sinkhorn algorithm.

14:00 **Fundamental limits for learning hidden Markov model parameters**
Kweku Abraham, Elisabeth Gassiat, Zacharie Naulet

Abstract: The densities making up a nonparametric mixture distribution, absent assumptions, cannot be identified from independent samples from the mixture. One approach to address this issue is to place assumptions on the distributions themselves, for example by limiting to Gaussian mixtures, or by assuming some form of separation for the different components of the mixture. An alternative approach is to introduce dependence in the sampling. Arguably the simplest dependence structure for a sequence of variables is a Markov chain. When the (unobserved) class labels in a mixture model follow a Markov chain, the observations are said to form a hidden Markov model. It is known, remarkably, that the model parameters (that is, the transition matrix of the underlying Markov chain and the densities of the data conditional on the possible values of the chain) are identifiable under virtually no conditions. In particular, when there are two

hidden states, it suffices to rule out degenerate sub-models in which the observations are independent. I will outline the three distinct ways in which the model can degenerate into the independent subcase, and give a quantitative notion of the "distance" from independence required (as a function of the number of observations) for model parameters to be learnable. I will highlight some implications of this learnability on clustering the data using an empirical Bayes classifier.

14:30 Deconvolution with general and unknown noise distribution
Luc Lehéricy, Sylvain Le Corff, Elisabeth Gassiat

Abstract: The objective of the deconvolution problem is to recover a signal based on the sum of this signal and an independent noise. Since the Fourier transform of the distribution of such observations is the product of the Fourier transforms of the signal and of the noise distributions, the standard approach is to estimate the Fourier transform of the observation, divide by the Fourier transform of the noise, and invert the Fourier transform. Crucially, this approach requires to know the noise distribution, or estimate it using a sample of pure noise, and isn't usable when the noise distribution is unknown. We consider the deconvolution problem for multidimensional signals when no information is available on the noise distribution. We show that the distribution of both the signal and the noise can be recovered based solely on the distribution of the observations, provided that the signal has a Laplace transform with an exponential growth smaller than 2 and that it can be decomposed into two dependent components. The only assumption on the noise is that its two components are independent. Based on this result, we propose an estimator of the distribution of the signal and establish its rates of convergence when the signal ranges from compactly supported to having tails almost as heavy as Gaussian. The estimator is adaptive in the tail heaviness, and these rates are optimal when the signal is compactly supported.

Stochastic networks, graphs, and random matrices

Organiser: Anirban Dasgupta

Chair: Eugen Pircalabelu

Room: Leda

13:30 **Efficiency Lower Bounds for Distribution-Free Nonparametric Tests Based on Optimal Transport**

Nabarun Deb, *Bhaswar Bhattacharya*, Bodhisattva Sen

Abstract: The Wilcoxon rank-sum/Mann-Whitney test is one of the most popular distribution-free procedures for testing the equality of two univariate probability distributions. One of the main reasons for its popularity can be attributed to the remarkable result of Hodges and Lehmann (1956), which shows that the asymptotic relative efficiency of Wilcoxon's test with respect to Student's t-test, under location alternatives, never falls below 0.864, despite the former being exactly distribution-free in finite samples. Even more striking is the result of Chernoff and Savage (1958), which shows that the efficiency of a Gaussian score transformed Wilcoxon's test, against the t-test, is lower bounded by 1. In this talk we will discuss multivariate versions of these celebrated results, by considering distribution-free analogues of the Hotelling T^2-test based on optimal transport. The proposed tests are consistent against a general class of alternatives and satisfy Hodges-Lehmann and Chernoff-Savage-type efficiency lower bounds over various natural families of multivariate distributions, despite being entirely agnostic to the underlying data generating mechanism. Analogous results for independence testing will also be discussed.

14:00 Network Regression and Supervised Centrality Estimation
Junhui Cai, Dan Yang, Wu Zhu, Haipeng Shen, *Linda Zhao*

Abstract: The centrality in a network is often used to measure nodes' importance and model network effects. In empirical studies, a two-stage procedure, which first estimates the centrality and then infers the network effect from the estimated centrality, is widely adopted, but lacks theoretical understanding. We propose a unified framework, under which we prove the two-stage's shortcomings in centrality estimation and network effect inference. Furthermore, we propose a supervised centrality estimation methodology, whose advantages in both estimation and inference are proved theoretically and demonstrated via extensive simulation and a case study in predicting currency risk premiums from the global trade network. This is a joint work with Junhui Cai, Haipeng Shen, Dan Yang, and Wu Zhu.

14:30 Community detection on probabilistic graphical models with group-based penalties
Gerda Claeskens, *Eugen Pircalabelu*

Abstract: A new strategy of probabilistic graphical modeling is developed that draws parallels from social network analysis. Probabilistic graphical modeling summarizes the information coming from multivariate data in a graphical format where nodes, corresponding to random variables, are linked by edges that indicate dependence relations between the nodes. The purpose is to estimate the structure of the graph (which nodes connect to which other nodes) when data at the nodes are available. On the opposite side of the spectrum, social network analysis considers the graph as the observed data. Given thus the graph where connections between nodes are observed rather than estimated, social network analysis estimates models that represent well an underlying mechanism which has generated the observed graph. We propose a new method that exploits the strong points of each framework as it estimates jointly an undirected graph and communities of homogenous nodes, such that the structure of the communities is taken into account when estimating the graph and conversely, the structure of the graph is accounted for when estimating homogeneous communities of nodes. The procedure uses a joint group graphical lasso approach with community detection-based grouping, such that some groups of edges co-occur in the estimated graph. The grouping structure is unknown and is estimated based on community detection algorithms. Theoretical derivations regarding graph convergence and sparsistency, as well as accuracy of community recovery are included, while the method's empirical performance is illustrated in an fMRI context, as well as with simulated examples.

Recent advances in dimension reduction and functional data analysis

Organiser: Bing Li

Chair: Efstathia Bura

Room: Athena

13:30 Statistical Inference For Functional Linear Quantile Regression
Peijun Sang

Abstract: We propose inferential tools for functional linear quantile regression where the conditional quantile of a scalar response is assumed to be a linear functional of a functional covariate. In contrast to conventional approaches, we employ kernel convolution to smooth the original loss function. The coefficient function is estimated under a reproducing kernel Hilbert space framework. A gradient descent algorithm is designed to minimize the smoothed loss function with a roughness penalty. With the aid of the Banach fixed-point theorem, we show the existence and uniqueness of our proposed estimator as the minimizer of the regularized loss function in an appropriate Hilbert space. Furthermore, we establish the convergence rate as well as the weak convergence of our estimator. As far as we know, this is the first weak convergence result for a functional quantile regression model. Pointwise confidence intervals and a simultaneous confidence band for the true coefficient function are then developed based on these theoretical properties. Numerical studies including both simulations and a data application are conducted to investigate the performance of our estimator and inference tools in finite sample.

14:00 Nonparametric graphical models for high-dimensional functional data
Eftychia Solea, Holger Dette

Abstract: We consider the problem of constructing nonparametric undirected graphical models for high- dimensional functional data. Most existing statistical methods on graphical models assume either the Gaussian distribution on the vertices or linear conditional means. In this article we provide a more flexible model which relaxes the linearity assumption by replacing it by an arbitrary additive form. The utilisation of the functional principal components offers an estimation strategy that uses a group lasso penalty to estimate the relevant edges of the graph. We establish the model selection consistency for the resulting estimator, while allowing both the number of predictors and the number of functional principal components to diverge to infinity with increasing sample size. We investigate the empirical performance of our method through simulation studies and a real data application.

14:30 Functional Single-Index Models and Gaussian Stein's Identity
Bharath Sriperumbudur, Krishna Balasubramanian, Hans-Georg Mueller

Abstract: In this work, we consider the estimation of the parametric component of a functional single-index model that relates the functional covariate, which is a sample-path of a zero-mean Gaussian process, to a scalar-valued response. Based on the infinite-dimensional extension of Gaussian Stein's identity, which allows estimating the parametric component while being oblivious to the non-parametric component, we propose an estimator that involves solving a ridge regression problem in a reproducing kernel Hilbert space (RKHS). We establish convergence rates for this estimator in the regimes of commutativity and non-commutativity of the kernel integral operator and the covariance operator of the Gaussian process. These results also recover the optimal rates for prediction error when the model is linear and the operators commute.

15:00 Sufficient dimension reductions for mixed predictors
Efstathia Bura, Liliana Forzani, Pamela Llop, Rodrigo Garcia Arancibia, Diego Tomassi

Abstract: Most data sets comprise of measurements on continuous and categorical variables. Yet, modeling high-dimensional mixed predictors has received limited attention in regression and classification Statistics literature. We study the general regression problem of inferring on a variable of interest based on high dimensional mixed continuous and binary predictors. The aim is to find a lower dimensional function of the mixed predictor vector that contains all the modeling information in the mixed predictors for the response, which can be either continuous or categorical. The approach we propose identifies sufficient reductions by reversing the regression and modeling the mixed predictors conditional on the response. We derive the maximum likelihood estimator of the sufficient reductions, asymptotic tests for dimension, and a regularized estimator, which simultaneously achieves variable (feature) selection and dimension reduction (feature extraction). We study the performance of the proposed method and compare it with other approaches through simulations and real data examples.

ISNPS 2022

Venue Hotel

CORAL BEACH RESORT - 5*

GENERAL INFORMATION

The Coral Beach is a 5 star hotel situated on 300 metres of natural sandy beach with its own private harbour and boat. It is 17 minutes from Paphos city centre, 35 minutes from Paphos International Airport and 1 hour and 30 minutes from Larnaca International Airport. Adjacent to the Akamas peninsula, an area protected by UNESCO, the Coral Beach Hotel & Resort offers the perfect base for exploration.

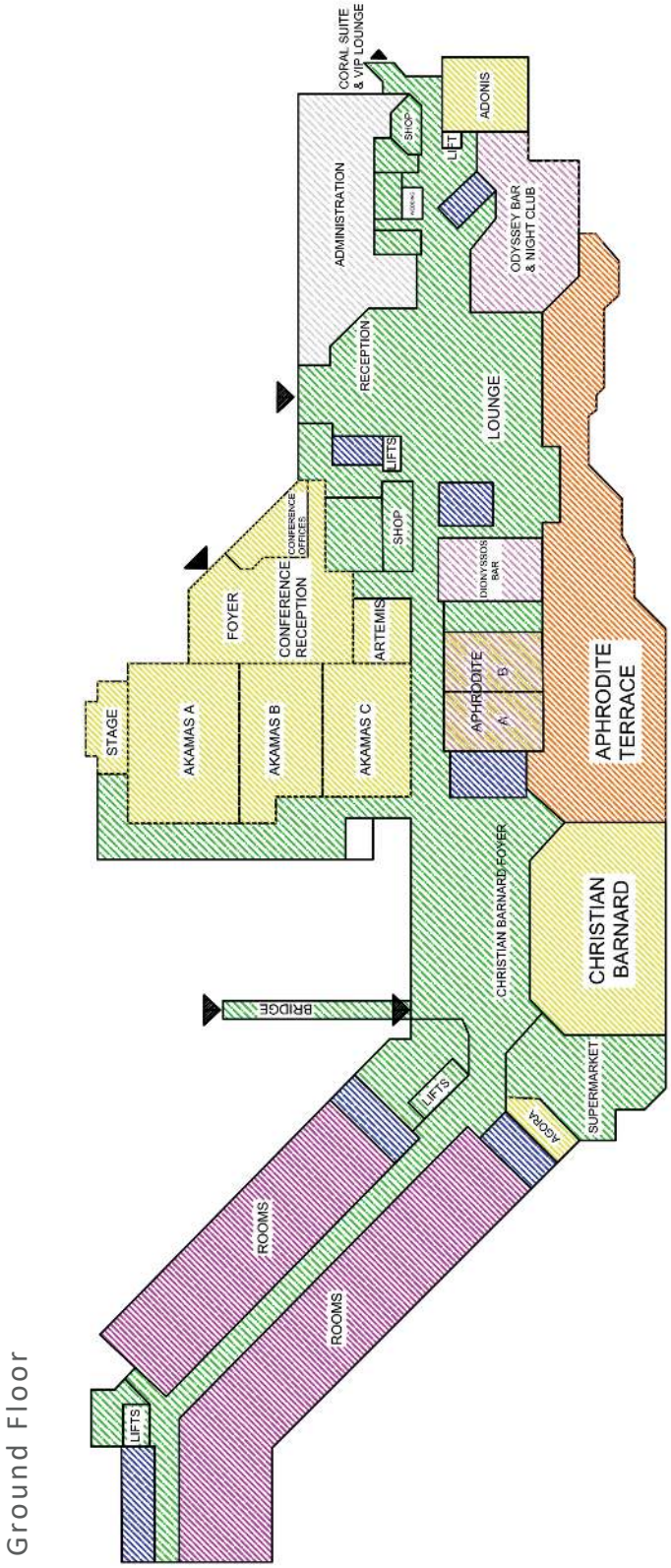
This unique resort combines the traditional Cypriot decor of white walls and authentic woodwork with the modern amenities expected of a five star resort.

Contact Details

Address: Coral Bay Ave, Coral Bay,
Paphos 8099

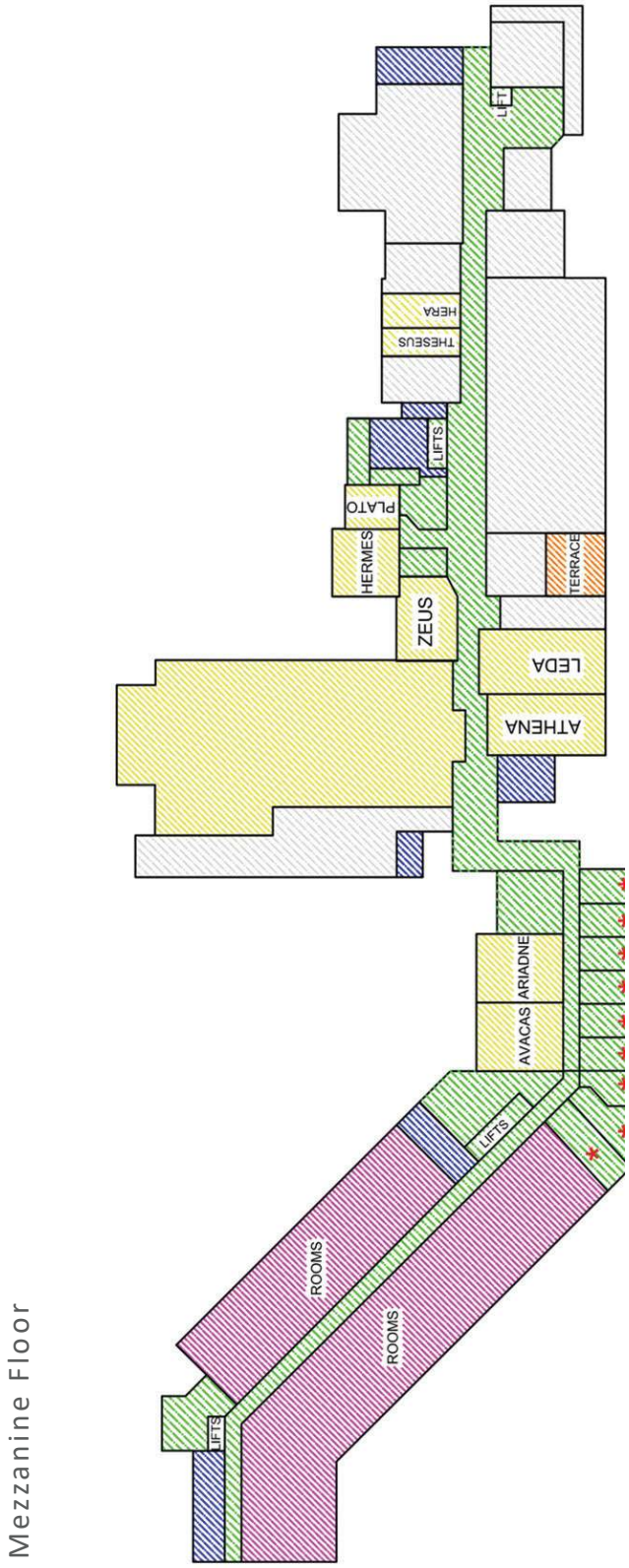
Tel: +357 26 88 1000
E-mail: info@coral.com.cy

VENUE FLOORPLAN



Ground Floor

VENUE FLOORPLAN



Mezzanine Floor

SOCIAL EVENTS

WELCOME COCKTAIL

Date: **20 June, 2022**
Time: **19:00 – 20:30**
Where: Venue (In the hotel grounds)

Welcome Cocktail is the first social gathering between all conference delegates and it will take place at the Venue Hotel. It will be a relaxing evening during which delegates will have the opportunity to talk to colleagues and peers, while enjoying local drinks and ample canapés.
The Welcome Cocktail is included in all Registration Fees

Ticket per accompanying person: **€35.00**

TOUR & CONFERENCE DINNER

Date: **22 June 2022**
Time: **16:00- 23:00**
Departure Time: **16:00**
Departure From: **Venue Hotel Lobby**

We will get together at the lobby of the Venue Hotel, from where we will promptly depart in air-conditioned coaches for a city tour. A professional guide will tell us about the history of Cyprus and Paphos Town in particular. Dinner will take place at a traditional tavern starting at 20:30 serving excellent dishes of Cypriot cuisine complimented with local drinks and desserts.

The Conference Dinner is included in all Registration Fees.

Ticket per accompanying person or fees participant: **€60.00**

PAPHOS

Paphos, a city rich in history and culture, is the gem of western Cyprus. Believed to be the birthplace of Aphrodite, the Greek goddess of love and beauty, Paphos proudly boasts the remains of palaces, theatres, fortresses and tombs that belong to Classical, Hellenistic and Roman periods. There is also archaeological evidence supporting the city's existence from the Neolithic period.

All these elements and facts give Paphos a remarkable architectural and historical value, and this is mostly why the town of Paphos with the Mosaics palaces and Tombs of the Kings, is included in the official UNESCO list of cultural and natural treasures of the world's heritage.

Paphos was valued as a major port and the capital of Cyprus during Roman times. Today, this small harbour with a population of about 32,754, has slowly and steadily emerged as an attractive, popular tourist destination. Ktima is the main residential district while Kato Paphos, by the sea, is built around the medieval castle and contains most of the luxury hotels and the entertainment infrastructure of the city. Hundreds of shops, restaurants, bars and a newly built shopping mall complete the picture of this exceptional town.

In the district of Paphos and within a half an hour drive away, one can visit numerous picturesque villages with traditional tavernas, churches and archaeological sites. Polis, Akamas Peninsula, Aphrodite's Rock, Peyeia, Argaka, Lara Bay with the Caretta- Caretta turtles are just a few to name. Paphos, along with Aarhus, Denmark, were the European Capitals of Culture in 2017.

ABOUT PAPHOS


Medieval Paphos Castle


Sunset


Petra tou Romiou


Tombs of the Kings


Coral bay


Ancient Odeon


Akamas Peninsula


Harbour promenade

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

FLEXIBLE SOLUTIONS TO SUIT YOUR SPECIFIC NEEDS

Easy Conferences Ltd has been in business since 1992 and has been specializing in the complete coordination and organization of conferences and all related activities. Through the development of its own online registration software, the company has expanded its operations outside Cyprus. We have extensive experience in organizing events ranging from 20 to 2000 participants for physical, hybrid or online participation. We consult, manage and assist in every step of the process of any event and we deliver top professional services throughout.

Our services extend from digital support, media promotion, conference website development and management, to the management of all conference related activities, complete interaction with suppliers and participants, online/onsite registration with secretariat, technical equipment and 24/7 phone help line. We are adaptable and extremely flexible as we are aware of the unique requirements and budget restrictions of each conference. Our services may be provided on an all-inclusive or on an a-la-carte basis.

Easy Conferences can provide organizers with a complete paper submission evaluation system at www.easyacademia.org. We also have our own, custom-made one-stop-shop Conference Management System, www.easyconferences.org which offers participants the ability to sign up and withi minutes register for the conference and its extra activiti participants accommodation, airport transfers, social for themselves and their accompanied persons and s their pay instantly online.

Our extensive experience and personal attention to e participant's needs, backed by our team members' u expertise in their field, as well as the selection of the r partners, has resulted in our impeccable track record that is our guarantee for perfectly organizing any conference or event.

Please visit our website, www.easyconferences.eu for more information on our services, a list of upcoming and past events, as well as referrals from our customers.

P.O.Box 24120, 1701, Nicosia, Cyprus
+357 22 591900
+357 22 591700
info@easyconferences.eu

www.easyconferences.eu
www.easyconferences.org

Bernoulli Society
for Mathematical Statistics
and Probability

Coordinator: