Department of Statistics Archive of Statistics Seminars - Michaelmas Term 2022

Friday 14 October, 3-4pm - Kean Ming Tan (University of Michigan)

This event took place in the Leverhulme Library COL 6.15.

Title - Convolution-Type Smoothing Approach for Quantile Regression

Abstract - Quantile regression is a powerful tool for learning the relationship between a response variable and a multivariate predictor while exploring heterogeneous effects. However, the non-smooth piecewise linear loss function introduces challenges to the computational aspect when the number of covariates is large. To address the aforementioned challenge, we propose a convolution-type smoothing approach that turns the non-differentiable quantile piecewise linear loss function into a twice- differentiable, globally convex, and locally strongly convex surrogate, which admits a fast and scalable gradient-based algorithm to perform optimization. In the low-dimensional setting, we establish nonasymptotic error bounds for the resulting smoothed estimator. In the high-dimensional setting, we propose the concave regularized smoothed quantile regression estimator, which we solve using a multi-stage convex relaxation algorithm. Theoretically, we characterize both the algorithmic error due to non-convexity and statistical error for the resulting estimator simultaneously. We show that running the multi-stage algorithm for a few iterations will yield an estimator that achieves the oracle property. Our results suggest that the smoothing approach leads to a significant computational gain without a loss in statistical accuracy.

Biography - Kean Ming Tan is currently an assistant professor at the Department of Statistics at University of Michigan. Previously, he was an assistant professor at the School of Statistics at University of Minnesota, and a postdoctoral research associate supervised by Han Liu and Tong Zhang. He joined the University of Washington in 2011 for his PhD degree, under the supervision of Daniela Witten. Kean is a statistician working on statistical machine learning methods for analyzing complex data sets. He develops multivariate statistical methods such as probabilistic graphical models, cluster analysis, discriminant analysis, and dimension reduction to uncover patterns from massive data set. He also works on topics related to robust statistics, quantile regression, non-convex optimization, and data integration from multiple sources. More recently, he is involved in applying instrumental variable to models with unmeasured confounders.

……..

Friday 21 October, 3-4pm - David Kaplan (University of Wisconsin)

This event took place in the Leverhulme Library COL 6.15.

Title - Bayesian Methods for Borrowing Historical Information With Applications to the Analysis of Large-Scale Assessments.

Biography - David Kaplan is the Patricia Busk Professor of Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin – Madison. Dr. Kaplan holds affiliate appointments in the University of Wisconsin's Department of Population Health Sciences and the Center for Demography and Ecology. Dr. Kaplan's program of research focuses on the development of Bayesian statistical methods for education research. His work on these topics is directed toward applications to large-scale cross-sectional and longitudinal survey designs. Dr. Kaplan is an elected member of the National Academy of Education and serves as the chair of its Research Advisory Committee; a recipient of the Samuel J. Messick Distinguished Scientific Contributions Award from the American Psychological Association (Division 5); a past-President of the Society for Multivariate Experimental Psychology; a fellow of the American Psychological Association (Division 5); a recipient of the Alexander Von Humboldt Research Award; an Honorary Research Fellow in the Department of Education at the University of Oxford., a fellow of the Leibniz Institute for Educational Research and Information and the Leibniz Institute for Educational Trajectories; and was a Jeanne Griffith Fellow at the National Center for Education Statistics. Dr. Kaplan received his Ph.D. in education from UCLA in 1987.

…………………………………………………………………………………………..

Monday 24 October, 3-4pm - Jianqing Fan (Princeton University)

This event took place in the Leverhulme Library COL 6.15.

Title - Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression.

Abstract - We introduce a Factor Augmented Sparse Throughput (FAST) model that utilizes both latent factors and sparse idiosyncratic components for nonparametric regression. The FAST model bridges factor models on one end and sparse nonparametric models on the other end. It encompasses structured nonparametric models such as factor augmented additive model and sparse low-dimensional nonparametric interaction models and covers the cases where the covariates do not admit factor structures. Via diversified projections as estimation of latent factor space, we employ truncated deep ReLU networks to nonparametric factor regression without regularization and to more general FAST model using nonconvex regularization, resulting in factor augmented regression using neural network (FARNN) and FAST-NN estimators respectively. We show that FAR-NN and FAST-NN estimators adapt to unknown low-dimensional structure using hierarchical composition models in nonasymptotic minimax rates. We also study statistical learning for the factor augmented sparse additive model using a more specific neural network architecture. Our results are applicable to the weak dependent cases without factor structures. In proving the main technical result for FAST-NN, we establish new a deep ReLU network approximation result that contributes to the foundation of neural network theory. Our theory and methods are further supported by simulation studies and an application to macroeconomic data. (Joint work with Yihong Gu)

Biography - Jianqing Fan, is a statistician, financial econometrician, and data scientist. He is Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at the Princeton University where he chaired the department from 2012 to 2015. He is the winner of The 2000 COPSS Presidents' Award, Morningside Gold Medal for Applied Mathematics (2007), Guggenheim Fellow (2009), Pao-Lu Hsu Prize (2013) and Guy Medal in Silver (2014). He got elected to Academician from Academia Sinica in 2012.

…………………………………………………………………………………………..

Monday 31 October, 3.15-4.15pm - Esther Ruiz Ortega (Universidad Carlos III de Madrid)

This event took place in the Leverhulme Library COL 6.15.

Title - Modelling and forecasting intervals of minimum/maximum temperature in the Iberian Peninsula.

Abstract - In this paper, we propose a novel methodology to model and forecast intervals of minimum and maximum temperature based on fitting state space models to center and log-range temperature. In doing so, we allow the center and log-range temperature to be related and to obtain measures of the uncertainty associated with estimates of the temperature trend and dispersion. The methodology is first implemented separately to intervals of minimum and maximum temperature observed monthly in four locations in the Iberian Peninsula chosen to represent different climate conditions. Namely, we consider temperatures in Barcelona, Coruña, Madrid and Seville. Second, given that, at each location, center and logrange temperature are shown to be unrelated, we fit a multivariate dynamic factor model to extract potential commonalities among center (log-range) temperature observed at a large number of locations in the Iberian Peninsula.

Biography - As of January 2020, the database RePEc/IDEAS places Esther Ruiz ORtega in the top 2.4% among female world economists (352 out of 14683) and in the top 2.4% among economists in Spain (56 out of 2310). Furthermore, she is in the top 5% authors worldwide according to several criteria as, for example, Number of Distinct Works, Number of Citations and Record of Graduates.

…………………………………………………………………………………….

<u>Friday 11 November, 3-4pm - Wolfgang Polonik (University of California)</u>

This event took place in the Leverhulme Library COL 6.15.

Title - Topologically penalized regression on manifolds.

Abstract - We study a regression problem on a compact manifold. In order to take advantage of the underlying geometry and topology of the data, we propose to perform the regression task on the basis of eigenfunctions of the Laplace-Beltrami operator of the manifold that are regularized with topological penalties. We will discuss the approach and the penalties, provide some supporting theory and illustrate the performance of the methodology on some data sets, illustrating the relevance of our approach in the case where the target function is ``topologically smooth". This is joint work with O. Hacquard, K. Balasubramanian, G. Blanchard and C. Levrard.

Biography - Wolfgang Polonik is a professor at the Department of Statistics, University of California, Davis. He received his Ph.D. degree from Ruprecht-KarlsUniversität Heidelberg in 1992. His areas of interest cover Nonparametric Statistics, Shape constraints, modality, Nonstationary Time series and Empirical process theory. Currently, he is specialized in Topological Data Analysis.

…………………………………………………………………………………………….

Department of Statistics Archive of Statistics Seminars – Lent Term 2023

<u>Friday 27 January, 2-3pm - Wei Zhong (Xiamen University)</u>

This event took place on Zoom.

Title - Semi-Distance Correlation and Its Applications.

Abstract - We propose a new measure of dependence between a categorical random variable and a random vector with potentially high dimensions, named semi-distance correlation. It is an interesting extension of distance correlation to accommodate the information of the categorical random variable. It equals zero if and only if the categorical random variable and the other random vector are independent. Two important applications of semi-distance correlation are considered. First, we develop a semi-distance independence test between a categorical random variable and a random vector and derive its asymptotic distributions. When the dimension of the random vector tends to infinity, we derive the explicit asymptotic normal distribution of the test statistic under the null hypothesis, which allows us to compute p-values in an efficient and fast way for high dimensional data. Second, we propose to use the semi-distance correlation as a marginal utility between the response and a group of covariates to do groupwise variable screening for ultrahigh dimensional classification problems. The sure screening property has also been established. Monte Carlo simulations and a real data application are presented to demonstrate the excellent finite sample property of the proposed procedures. A new R package semidist is also developed to implement the proposed methods.

Bio - Wei Zhong, Professor of Wang Yanan Institute for Studies in Economics (WISE) and School of Economics, and Department Chair, Department of Statistics and Data Science, School of Economics at Xiamen University. He obtained PhD degree in statistics from Pennsylvania State University in 2012. His main research interests include statistical learning for high dimensional data, hypothesis testing and econometrics. He has published more than 30 papers in *Annals of Statistics, Journal of the American Statistical Association, Biometrika, Journal of Econometrics, Journal of Business & Economic Statistics, Biometrics, Annals of Applied Statistics, Statistica Sinica* etc. He serves as an associate editor for Journal of the American Statistical Association (2023-present), Statistical Analysis and Data Mining (2018-present), The Canadian Journal of Statistics (2019-2021). He has received several grants including National Natural Science Foundation of China (Excellent Young Scholar Program, Key Project Program, General Program etc), National Statistical Science Research Grants of China, National Key R&D Program of China.

…………………………………………………………………………………….

Friday 10 February, 4-5pm - Annie Qu (University of California)

This event took place on Zoom.

Title - Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling.

Abstract - Crowdsourcing has emerged as an alternative solution for collecting large scale labels. However, the majority of recruited workers are not domain experts, so their contributed labels could be noisy. In this talk, we propose a two-stage model to predict the true labels for multicategory classification tasks in crowdsourcing. In the first stage, we fit the observed labels with a latent factor model and incorporate subgroup structures for both tasks and workers through a multi-centroid grouping penalty. Group-specific rotations are introduced to align workers with different task categories to solve multicategory crowdsourcing tasks. In the second stage, we propose a concordance-based approach to identify high-quality worker subgroups who are relied upon to assign labels to tasks. In theory, we show the estimation consistency of the latent factors and the prediction consistency of the proposed method. The simulation studies show that the proposed method outperforms the existing competitive methods, assuming the subgroup structures within tasks and workers. We also demonstrate the application of the proposed method to real world problems and show its superiority.

Bio - Qu's research focuses on solving fundamental issues regarding structured and unstructured large-scale data, and developing cutting-edge statistical methods and theory in machine learning and algorithms on personalized medicine, text mining, recommender systems, medical imaging data and network data analyses for complex heterogeneous data. The newly developed methods are able to extract essential and relevant information from large volume high-dimensional data. Her research has impacts in many fields such as biomedical studies, genomic research, public health research, social and political sciences.

Before she joins the UC Irvine, Dr. Qu is Data Science Founder Professor of Statistics, and the Director of the Illinois Statistics Office at the University of Illinois at Urbana-Champaign. She was awarded as Brad and Karen Smith Professorial Scholar by the College of LAS at UIUC, a recipient of the NSF Career award in 2004-2009. She is a Fellow of the Institute of Mathematical Statistics, a Fellow of the American Statistical Association, and a Fellow of American Association for the Advancement of Science. She is also a recipient of Medallion Award and Lecturer in 2024. She is JASA Theory and Methods co-editor in 2023-2025.

…………………………………………………………………………………….

Friday 24 February, 2-3pm - Anton Rask Lundbord (University of Copenhagen)

This event took place in 32 Lincoln's Inn Fields (32L.1.05).

Title - Modern Methods for Variable Significance Testing.

Abstract - Testing the significance of a variable or group of variables X for predicting a response Y, given additional covariates Z, is a ubiquitous task in statistics. A simple but common approach is to specify a linear model, and then test whether the regression coefficient for X is non-zero. However, when the model is misspecified, the test may have poor power, for example when X is involved in complex interactions, or lead to many false rejections. In this talk we study the problem of testing the model-free null of conditional mean independence, i.e. that the conditional mean of Y given X and Z does not depend on X. We discuss two recent proposals, one for real-valued Y with a focus on maximising power and another specific to functional X and Y, that are both able to leverage flexible nonparametric or machine learning methods, such as additive models or random forests, to yield both robust error control and high power. The methods come with uniform asymptotic guarantees and numerical experiments demonstrate the effectiveness of the approaches both in terms of maintaining Type I error control, and power, compared to several existing approaches.

Bio - Dr. Anton Rask Lundbord is a postdoc at the Copenhagen Causality Lab, University of Copenhagen working under the supervision of Dr. Niklas Pfister. They are currently working on causal inference and semiparametric problems related to the analysis of compositional data with an application to microbiome data in mind. Previously, Anton was a PhD student with Prof. Rajen Shah and Prof. Richard Samworth at the University of Cambridge.

……………………………………………………………………………………………

Friday 10 March, 2-3pm - Guosheng Yin (Imperial College)

This event took place in 32 Lincoln's Inn Fields (32L.1.05).

Title - Deep Learning and Generative Model for Conditional Survival and Hazard Functions.

Abstract - We propose a deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We also propose a two-stage generative approach for estimating the conditional cumulative distribution function for current status data which are commonly encountered in modern medicine, econometrics and social science. We first learn a conditional generator nonparametrically for the joint conditional distribution of observation times and event status, and then construct the nonparametric maximum likelihood estimators of conditional distribution functions based on samples from the conditional generator. Subsequently, we study the convergence properties of the proposed estimator and establish its consistency. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

Joint work with Wen Su, Changyu Liu, Kin-Yat Liu, Xingqiu Zhao and Jian Huang

Bio - Guosheng Yin is currently a Chair Professor in Statistics in Department of Mathematics at Imperial College London. Previously, he was Patrick SC Poon Endowed Professor and also served as the Head of Department of Statistics and Actuarial Science at The University of Hong Kong (2017-2023). Before that, he worked in Department of Biostatistics at University of Texas M.D. Anderson Cancer Center (2003-2009) after he received Ph.D. in Biostatistics from University of North Carolina at Chapel Hill in 2003. He was elected as a Fellow of the Institute of Mathematical Statistics in 2021, a Fellow of the American Statistical Association in 2013. He served as Associate Editor for Journal of American Statistical Association, Bayesian Analysis, Statistical Analysis and Data Mining, and Deputy Editor for Contemporary Clinical Trials. His main research areas include clinical trial methodology, adaptive design, Bayesian methods, survival analysis, high-dimensional data analysis, machine learning and AI. He has published more than 200 peer-reviewed papers in statistical and medical journals and AI or machine learning conferences as well as two books on clinical trial design and adaptive methods.

…………………………………………………………………

Friday 17 March, 2.30-3.30pm - Regina Liu (Rutgers University)

This event took place in 32 Lincoln's Inn Fields (32L.1.05).

Title - Fusion Learning: Fusion and i-Fusion (individualized Fusion) Learning.

Abstract - Advanced data collection technology nowadays has often made inferences from diverse data sources easily accessible. Fusion learning refers to combining inferences from multiple sources or studies to make a more effective overall inference than that from any individual source or study alone. We focus on the tasks: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently? 3) How to combine inference to enhance an individual or target study?

We present a general framework for nonparametric and efficient fusion learning for inference on multi-parameters, which may be correlated. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), which is developed by combining data depth, bootstrap and confidence distributions. We show that a depth-CD is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing inferential tool. The approach is shown to be efficient, general and robust. It readily applies to heterogeneous studies with a broad range of complex and irregular settings. This property also enables the approach to utilize indirect evidence from incomplete studies to gain efficiency for the overall inference. The approach will be shown with simulation studies and real applications in aircraft landing performance tracking and in financial forecasting.

This talk contains joint works with Dungan Liu (University of Cincinnati), Jieli Shen (Goldman Sachs) and Minge Xie (Rutgers University).

Bio - Regina Liu received her PhD in statistics from Columbia University and is currently Distinguished Professor of Statistics at Rutgers University. Her research areas include data depth, resampling, confidence distribution, and fusion learning. Aside from theoretical and methodological research, she has long collaborated with the FAA on aviation safety research projects on process control, text mining and risk management. She has served as Editor for the JASA and the Journal of Multivariate Analysis, and as Associate Editor for several journals, including the Annals of

Statistics. She is an elected fellow of the MS and the ASA. Among other distinctions, she is the recipient of 2021 ASA Noether Distinguished Scholar Award, and 2011 Stieltjes Professorship from Thomas Stieltjes Institute for Mathematics, The Netherlands, She has delivered an IMS Medallion Lecture among other named lectures. She was elected President of the Institute of Mathematical Statistics, 2020-2021.

…………………………………………………………………………….

Department of Statistics Archive of Statistics Seminars – Summer Term 2023

Friday 26 May, 2-3pm - Ruggero Bellio (University of Udine)

This event took place in Parish Hall (PAR.1.02).

Title - Scalable Estimation of Probit Models with Crossed Random Effects.

Abstract - This talk illustrates a scalable approach to mixed effects modeling with a probit link and a crossed random effects error structure. Random effects with a crossed structure arise often in social and business applications, a notable setting being that of electronic commerce, with random effects related to customers and purchased items, respectively. In sparsely sampled crossed data the computation for both frequentist and Bayesian estimation can easily grow superlinearly with respect to the sample size, which severely limits the use of these models for very large settings. The proposed method belongs to the class of composite likelihood estimators, and entails the fit of three misspecified reduced models. The resulting estimator is consistent and has an overall computational cost linear in the number of observations. This is a joint work with Art Owen and Swarnadip Ghosh, Stanford University, and Cristiano Varin, Ca’ Foscari University of Venice.

Bio - Ruggero Bellio is Professor of Statistics at the Department Of Economics and Statistics of the University Of Udine. He received the PhD In Statistics from the University of Padova in 2000. His research interests are likelihood methods, mixed models and their applications in various fields, statistical computing, data mining. He is very keen on the R statistical software. He served as Co-Editor (2016-2019) of the journal Statistical Methods and Applications.

…………………………………………………………….

Friday 2 June, 12-1pm - Yao Xie (Georgia Institute of Technology)

This event took place in Parish Hall (PAR.2.03).

Title - Conformal prediction for time series.

Abstract - We develop a general framework for constructing distribution-free prediction intervals for time series. Theoretically, we establish explicit bounds on conditional and marginal coverage gaps of estimated prediction intervals, asymptotically converging to zero under additional assumptions. We obtain similar bounds on the size of set differences between oracle and estimated prediction intervals. Methodologically, we introduce computationally efficient EnbPI and SPCI algorithms that wrap around ensemble predictors closely related to standard conformal prediction (CP) but do not require data exchangeability. Our algorithms avoid data-splitting and are computationally efficient by avoiding retraining and thus scalable to produce prediction intervals sequentially. We perform extensive simulation and real-data analyses to demonstrate its effectiveness compared with existing methods.

Bio - Yao Xie is the Coca-Cola Foundation Chair and Professor at Georgia Institute of Technology in the H. Milton Stewart School of Industrial and Systems Engineering and Associate Director of the Machine Learning Center. From September 2017 until May 2023 she was the Harold R. and Mary Anne Nash Early Career Professor. She received her Ph.D. in Electrical Engineering (minor in Mathematics) from Stanford University in 2012 and was a Research Scientist at Duke University. Her research lies at the intersection of statistics, machine learning, and optimization in providing theoretical guarantees and developing computationally efficient and statistically powerful methods for problems motivated by real-world applications. She received the National Science Foundation (NSF) CAREER Award in 2017, INFORMS Wagner Prize Finalist in 2021, and the INFORMS Gaver Early Career Award for Excellence in Operations Research in 2022. She is currently an Associate Editor for IEEE Transactions on Information Theory, IEEE Transactions on Signal Processing, Journal of the American Statistical Association, Theory and Methods, Sequential Analysis: Design Methods and Applications, INFORMS Journal on Data Science, and an Area Chair of NeurIPS and ICML.

…………………………………………………………………….

Friday 16 June, 12-1pm - Arnaud Doucet (University of Oxford)

This event took place in the Leverhulme Library (COL.6.15).

Title - From denoising diffusion models to diffusion Schrodinger bridges: generative modelling and inference.

Abstract - Denoising diffusion models are a novel powerful class oftechniques for generative modelling and inference at the core of popular applications such as Stable Diffusions, Dalle-2 or Midjourney. We will introduce these methods and present some of their limitations. We will then discuss how recent alternative based on transport ideas can resolve some of these limitations. In particular, we will focus on diffusion Schrodinger bridge, an entropy-regularized version of optimal transport, and discusshow it can be approximated numerically.

Bio - From 1 April 2023, Arnaud is a full-time Senior Research Scientist at DeepMind. He was an Institute of Mathematical Statistics Medallion (IMS) Lecturer in 2016, was elected IMS Fellow in 2017 and awarded the Guy Silver Medal from the Royal Statistical Society in 2020.

……………………………………………………………….

Tuesday 29 August, 2-3pm - Chun-houh Chen (Academia Sinica)

This event took place in the Leverhulme Library (COL.6.15).

Title - Matrix Visualization for Exploratory Data Analysis.

Abstract - "It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it" (Exploratory Data Analysis: John Tukey, 1977). Modern statisticians and data scientists are confronted with the challenge of deciphering increasingly high-dimensional and complex data, as traditional graphics and visualization tools fall short in meeting these demands. The ability to comprehend the overarching structure in massive datasets is an even more formidable task. Therefore, effective Exploratory Data Analysis (EDA) approaches, coupled with intuitive and user-friendly data visualization environments, will become increasingly vital in determining what can be achieved in the era of big data.

Matrix Visualization (MV) has proven to be more effective than standard EDA tools, such as Boxplot, Scatterplot (utilizing dimension reduction techniques), and Parallel Coordinate Plot, in extracting information from moderate to large (dimension and sample size) datasets. In this presentation, I will initially provide a concise overview of MV's technical foundation for continuous, binary, nominal, symbolic data, as well as data with cartographic links, using the Generalized Association Plots (GAP) developed by our information visualization laboratory. Subsequently, real-world applications addressing scientific issues in fields such as biological experiments, medical research, and social surveys will be showcased, followed by a discussion of ongoing advancements and potential future trajectories for MV research.

Bio - Website