Department of Statistics Archive of Data Science Seminars - Michaelmas Term 2022

Monday 17 October 2022, 2-3pm - Caroline Uhler (MIT)

Website

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

Title - From Interventions to Causality using Over-Parameterized Neural Networks.

Abstract - Massive data collection holds the promise of a better understanding of complex phenomena and ultimately, of better decisions. An exciting opportunity in this regard stems from the growing availability of perturbation / intervention data (for example from drug/knockout screens in biology, advertisement, online education, etc.). In order to obtain mechanistic insights from such data, a major challenge is the development of a framework that integrates observational and interventional data and allows causal transportability, i.e., predicting the effect of unseen interventions or transporting the effect of interventions observed in one context to another. I will discuss how over-parameterized neural networks can be used for these problems. In particular, I will characterize the implicit bias of over-parameterized autoencoders and link this to causal transportability in the context of virtual drug screening.

Biography - Caroline Uhler is a Full Professor in the Department of Electrical Engineering and Computer Science and the Institute for Data, Systems, and Society at MIT. In addition, she is a core institute member at the Broad Institute, where she co-directs the Eric and Wendy Schmidt Center. She holds an MSc in mathematics, a BSc in biology, and an MEd all from the University of Zurich. She obtained her PhD in statistics from UC Berkeley in 2011 and then spent three years as an assistant professor at IST Austria before joining MIT in 2015. She is a Simons Investigator, a Sloan Research Fellow, and an elected member of the International Statistical Institute. In addition, she received an NSF Career Award, a Sofja Kovalevskaja Award from the Humboldt Foundation, and a START Award from the Austrian Science Foundation. Her research lies at the intersection of machine learning, statistics, and genomics, with a particular focus on causal inference, representation learning, and gene regulation.

……

Monday 14 November 2022, 2-3pm - Vladimir Vovk (Royal Holloway, University of London)

Website

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

Title - Applications of e-values to multiple hypothesis testing.

Abstract - In this talk I will review two alternative tools for statistical hypothesis testing, p-values and e-values. Both have been used in the algorithmic theory of randomness for decades (on the log scale and under other names), but only p-values are widely used in non-Bayesian statistics; e-values are related to Bayes factors, especially in the case of a simple null hypothesis. The advantage of e-values is that they are easy to combine. This makes them a convenient and powerful tool for multiple hypothesis testing.

Biography - Vovk started working as a researcher in the Russian Academy of Sciences, then became a Fellow in the Center for Advanced Study in the Behavioral Sciences at Stanford University. He was appointed as a professor of Computer Science at Royal Holloway and Bedford New College, where he currently serves as co-director of the Centre for Machine Learning.

Early in his career, Vovk was heavily involved in the development of the foundations of probability, along with Glenn Shafer. Their work has resulted in a book, Probability and Finance: It's Only a Game!, published in 2001, which was subsequently translated into Japanese in 2006 by Masayuki Kumon and edited by Kei Takeuchi. In 2005, he co-invented the Conformal prediction framework with Alexander Gammerman.
Vovk has delivered speeches all around the world. In 2021, he was invited to deliver a series of memorial lectures to Prasanta Chandra Mahalanobis in India. On the 20-year anniversary of The Society for Imprecise Probability (SIPTA) in 2019, he was invited to deliver a talk on "Game-theoretic foundations for imprecise probabilities" in Belgium. In 2016, he delivered a seminar about "Probability-free theory of continuous martingales" at Imperial College in the UK. In 2014, he delivered a seminar at University of Hawai'i in the USA.

Vovk has written 9 books, more than 280 research papers, and has an estimated h-index of 53. He holds fellowship positions at Stanford University (USA), Arizona State University (USA) and Yandex (Russia).

…….

Monday 21 November 2022, 2-3pm - Anastasia Borovykh (Imperial College London)

Website

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

Title - Towards explainable and privacy-preserving machine learning.

Abstract - In the last decade, fuelled by drastic increases in computational power and the wide availability of data (i.e. big data), machine learning, and specifically deep neural networks, loosely inspired by neuronal structures in the brain, have been increasingly deployed in the real world. Despite the satisfactory performance achieved in practical applications, these models are generally difficult to analyse and their performance is not always fully understood. This impacts the deployment of neural network models as it directly influences two critical real-world challenges: generalisation - guaranteeing good performance of the model in unseen scenarios and privacy - ensuring the trained model does not give away sensitive information about the datasets it was trained on. In this talk we will first go into more detail on the challenges associated with generalisation and

privacy. We will then discuss several recent advancements in defining robust and privacy-preserving machine learning algorithms.

Biography - Anastasia Borovykh is currently an Assistant Professor (lecturer) at the Department of Mathematics at Imperial College London and the Imperial-X initiative. Her group works on computational models to understand and improve information processing in artificial and biological intelligent systems through a combination of tools from stochastic processes, statistical mechanics and mathematical modeling. They apply this in i) explainable machine learning, ii) privacy-preserving machine learning and iii) neuroscience.

Anastasia also enjoys collaborating with industry on applied problems that use machine learning and optimization in for example finance, smart cities and healthcare.
Prior to her current position, Anastasia was an Assistant Professor at the University of Warwick at the Operations Research department and did postdocs at Imperial College London and CWI Amsterdam. She obtained her PhD cum laude from the University of Bologna as part of a Marie-Curie ITN-EID project. Her MSc was in Quantitative Finance at the VU Amsterdam and my BSc in Applied Mathematics from the Delft University of Technology.

Department of Statistics Archive of Data Science Seminars - Lent Term 2023

Monday 30 January, 2-3pm - David Ginsbourger (University of Bern)

Website

This event took place in the Graham Wallas Room (OLD 5.25).

Title - On Gaussian Process multiple-fold cross-validation.

Abstract - In this talk I will give an overview of some recent results pertaining to the fast calculation of Gaussian Process multiple-fold cross-validation residuals and their covariances, as well as to kernel hyperparameter estimation via related approaches. At first, the focus will be put on results from (arXiv:2101.03108, joint work with Cedric Schärer), where fast Gaussian process leave-one-out formulae are generalized to multiple-fold cross-validation. A special focus will be put on the impact of designing the folds on covariance hyperparameter fitting. In particular, I will present results of a joint work with Athénaïs Gautier and Cédric Travelletti on an inverse problem from geosciences where considered formulae and criteria are applied to linear forms in the underlying GP and the way of partitioning observations is found to substantially affect range estimation.

Biography - David Ginsbourger is Professor (Extraordinarius) of Statistical Data Science and co-Director of the Institute of Mathematical Statistics and Actuarial Science at the University of Bern, where he is currently serving as Director of Studies in Statistics and leading the "Uncertainty Quantification and Spatial Statistics" research group. From 2015 to 2020, he mainly worked as a permanent senior researcher at Idiap Research Institute. He defended his PhD in Applied Mathematics at the Ecole Nationale Supérieure des Mines de Saint-Etienne in 2009. He is currently on the editorial boards of the SIAM/ASA Journal on Uncertainty Quantification and of Technometrics, as well as area chair/metareviewer at ICML 2022, NeurIPS 2022, and AISTATS 2023.

……

Monday 27 February, 2-3pm - Erwan Scornet (Ecole Polytechnique)

Website

This event took place in the Graham Wallas Room (OLD 5.25).

Title - Is interpolation benign for random forests?

Abstract - Statistical wisdom suggests that very complex models, interpolating training data, will be poor at predicting unseen examples. Yet, this aphorism has been recently challenged by the identification of benign overfitting regimes, specially studied in the case of parametric models: generalization capabilities may be preserved despite model high complexity. While it is widely known that fully-grown decision trees interpolate and, in turn, have bad predictive performances, the same behavior is yet to be analyzed for random forests. In this talk, I will present how the trade-off between interpolation and consistency takes place for several types of random forest models. In particular, I will establish that interpolation regimes and consistency cannot be achieved for non-adaptive random forests. Since adaptivity seems to be the cornerstone to bring together interpolation and consistency, we study the Median RF which is shown to be consistent even in the interpolation setting. Regarding Breiman's forest, we theoretically control the size of the interpolation area, which converges fast enough to zero, so that exact interpolation and consistency can occur in conjunction.

Biography - Since September 2016, Erwan Scornet is assistant professor at the

Center for Applied Mathematics (CMAP) in Ecole Polytechnique near Paris. His research interests focus on theoretical statistics and Machine Learning with a particular emphasis on nonparametric estimates. He did his PhD thesis on a particular algorithm of Machine Learning called random forests, under the supervision of Gérard Biau (LSTA - Paris 6) and Jean-Philipe Vert (Institut Curie).

…….

Monday 20 March, 2-3pm - Patrick Loiseau (INRIA, Ecole Polytechnique, ENSAE)

Website

This event took place in the Graham Wallas Room (OLD 5.25).

Title - Statistical discrimination in selection and matching.

Abstract - Discrimination in selection problems such as hiring or college admission is often explained by implicit bias of the decision-maker against a disadvantaged demographic group. In this talk, we argue that discrimination may occur from second-order statistical properties even in the absence of bias. We consider a model where the decision-maker receives a noisy estimate of each candidate's quality, whose variance depends on the demographic group of the candidate---we term this implicit (or differential) variance. We show that regardless of the information that the decision-maker has to make its selection (Bayesian or group-oblivious), differential variance leads to discrimination in the selection. We then study the effect of affirmative action policies on the selection quality and show that, in some cases, it may even increase the selection quality. Finally, we analyze a stable matching problem, where there are two decision-makers selecting from the same pool of candidates. We show that even in the absence of differential variance, a difference across groups in the correlation between the quality estimates of the two decision-makers leads to discrimination.

Biography - Patrick Loiseau is a researcher at Inria Saclay, and an adjunct Professor at Ecole Polytechnique and ENSAE (Palaiseau). He is the co-head of the FairPlay team, a joint team between Criteo, ENSEA, Ecole Polytechnique, and Inria. Since 2019, he is also the co-holder of a chair of the MIAI@Grenoble Alpes institute on “Explainable and Responsible AI”. Prior to joining Inria, he was an Assistant Professor of data science at EURECOM and he held long-term visiting positions at UC Berkeley and at the Max-Planck Institute for Software Systems (MPI-SWS) where

he was the recipient of a Humboldt fellowship for experienced researchers (2016). He works on game theory and machine learning, with a focus on societal and ethical aspects (fairness and privacy) and on security and privacy.

……

Monday 27 March, 4-5pm - Yeganeh Alimohammadi (Stanford)

Website

This event took place on Zoom.

Title - The Power of a Few Local Samples for Predicting Epidemics

Abstract - People's interaction networks play a critical role in epidemics. However, accurately mapping these interactions can be expensive and sometimes impossible, making it difficult to predict the likelihood and outcome of an outbreak. Instead, contact tracing a few samples from the population is enough to estimate an outbreak's likelihood and size. I will present a model-free estimator based on the contact tracing results and give theoretical guarantees on the estimator's accuracy for a large class of networks.

Biography - Yeganeh is a Ph.D. student in operations research at Stanford University, where she is advised by Amin Saberi. Her research interests are algorithm design and operations research with an emphasis on applications. In particular, she studies the theoretical grounds of network models of practical importance, mainly focusing on studying epidemics on networks, designing efficient sampling algorithms from large networks, and network optimization.

……

Department of Statistics Archive of Data Science Seminars - Summer Term 2023

Monday 15 May, 2-3pm - Marta Blangiardo (Imperial College)

Website

This event took place in the Sir Arthur Lewis Building (SAL.1.05).

Title - Can we use spatio-temporal wastewater data to battle COVID-19?

Abstract - The utility of wastewater-based epidemiology as an early warning tool has been explored widely across the globe during the current COVID-19 pandemic. However, no attempt has previously been made to develop a model that predicts wastewater viral concentration at fine spatio-temporal resolutions covering an entire country, a necessary step towards using wastewater monitoring for the early detection of local outbreaks. In this talk I will first show how we can model the relationship between weekly viral concentration in wastewater at specific locations and a collection of covariates covering socio-demographics, land cover and virus-associated genomic characteristics. I will then discuss the potential for joint modelling of wastewater data and COVID-19 prevalence on a space-time resolved domain to improve the performance of public health surveillance systems.

Biography - Marta Blangiardo is a professor of Biostatistics in the Department of Epidemiology and Biostatistics at Imperial College London and leads the Biostatistics and Data Science theme of the MRC Centre for Environment and Health (https://environment-health.ac.uk/). She is one of the PIs of the Turing-RSS Health Data lab (https://www.turing.ac.uk/research/research-projects/turing-rss-health-data-lab) and has an academic honorary contract with the UK Health Security Agency. Her main interests are related on the methodological aspects of environmental exposure estimation and on spatial and spatio-temporal models for disease mapping and for risk assessment.

……

Monday 12 June, 2-3pm - Marco Scutari (IDSIA)

Website

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

Title - Achieving fairness with a simple ridge penalty.

Abstract - The adoption of machine learning in applications where it is crucial to ensure fairness and accountability has led to a large number of model proposals in the literature, largely formulated as optimisation problems with constraints reducing

or eliminating the effect of sensitive attributes on the response. While this approach is very flexible from a theoretical perspective, the resulting models are somewhat black-box in nature: very little can be said about their statistical properties, what are the best practices in their applied use, and how they can be extended to problems other than those they were originally designed for. Furthermore, the estimation of each model requires a bespoke implementation involving an appropriate solver which is less than desirable from a software engineering perspective. In this talk, I will take the opposite view that classical statistical models can be adapted to enforce fairness while preserving their well-known properties and the associated best practices in their applied use. I discuss how combining generalised linear models (GLMs) with penalised regression works into a *fair (generalised) ridge regression* model works very well for this purpose. Results from real-world data show this approach has competitive predictive accuracy, simple to implement, and suitable to a wide variety of applications.

Biography - Marco Scutari is a Senior Researcher at Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Switzerland. He has held positions in Statistics, Statistical Genetics and Machine Learning at UCL, University of Oxford and IDSIA since completing his Ph.D. in Statistics in 2011. His research focuses on the theory of Bayesian networks and their applications to biological and clinical data, as well as statistical computing and software engineering. Recently he has worked on introducing fairness in financial applications in a 3-years-long project with UBS Switzerland.