Department of Statistics Archive of 'Statistics and Data Science' Seminars 2024

Autumn Term 2024

Friday 4 October 2024, 2-3pm - Tom Everitt (Google DeepMind)

This event will take place in COL.1.06.

Title: Robust Agents Learn Causal World Models

Abstract: It has long been hypothesised that causal reasoning plays a fundamental role in robust and general intelligence. However, it is not known if agents must learn causal models in order to generalise to new domains, or if other inductive biases are sufficient. We answer this question, showing that any agent capable of satisfying a regret bound under a large set of distributional shifts must have learned an approximate causal model of the data generating process, which converges to the true causal model for optimal agents. We discuss the implications of this result for several research areas including transfer learning and causal inference.

Biography: Tom Everitt is a Staff Research Scientist at Google DeepMind leading the Causal Incentives Working Group. His work is on AGI Safety, i.e. how we can safely build and use highly intelligent AI. His PhD thesis, Towards Safe Artificial General Intelligence, is the first PhD thesis specifically devoted to this topic. Since then, he has been building towards a theory of alignment based on Pearlian causality.

……………………………………………………………………………………………………

Wednesday 9 October 2024, 2-3pm - Xinwei Shen (ETH Zürich)

This event will take place in COL.1.06.

Title: Distributional learning: from methodology to applications

Abstract: Estimating the full (conditional) distribution is crucial to many applications. However, existing methods such as quantile regression typically struggle with high-dimensional response variables. To this end, distributional learning models the target distribution via a generative model, which enables inference via sampling. In this talk, we introduce a distributional learning method called engression. We then demonstrate the applications of engression to several statistical problems including extrapolation in nonparametric regression, causal effect estimation, and dimension reduction, as well as scientific problems such as climate downscaling.

Biography: Xinwei Shen is a postdoctoral researcher at the Seminar for Statistics, ETH Zürich. She obtained her PhD in the Department of Mathematics at Hong Kong University of Science and Technology in 2022, and obtained a Bachelor of Science degree at Fudan University in 2018.

....

Friday 18 October 2024, 4-5pm - Sofia Olhede (EPFL)

This event will take place in COL.1.06.

Title: Spectral estimation for discovery of partial associations in spatial point processes and random fields

Abstract: Spatial variables can be observed in many different forms, such as regularly sampled random fields (lattice data), point processes, and randomly sampled spatial processes. Joint analysis of such collections of observations is clearly desirable, but complicated by the lack of an easily implementable analysis framework. It is well known that Fourier transforms provide such a framework, but its form has eluded data analysts. We formalize it by providing a multitaper analysis framework using coupled discrete and continuous data tapers, combined with the discrete Fourier transform for inference. Using this set of tools is important, as it forms the backbone for practical spectral analysis. In higher dimensions it is important not to be constrained to Cartesian product domains, and so we develop the methodology for spectral analysis using irregular domain data tapers, and the tapered discrete Fourier transform. We discuss its fast implementation, and the asymptotic as well as large finite domain properties. Estimators of partial association between different spatial processes are provided as are principled methods to determine their significance, and we demonstrate their practical utility on a large-scale ecological dataset. This is a joint work with Jake P. Grainger, Tuomas A. Rajala, David J. Murrell.

Biography: Sofia Olhede is a professor of Statistics at EPFL in Switzerland. She joined UCL prior to this in 2007, before which she was a senior lecturer of statistics (associate professor) at Imperial College London (2006-2007), a lecturer of statistics (assistant professor) (2002-2006), where she also completed her PhD in 2003 and MSci in 2000. She has held three research fellowships while at UCL: UK Engineering and Physical Sciences Springboard fellowship as well as a five-year Leadership fellowship, and now holds a European Research Council Consolidator fellowship. Sofia has contributed to the study of stochastic processes; time series, random fields and networks. Sofia was part of the multi-institutional team that set up the UK national data science institute, the Alan Turing Institute. She organised and served as chair of the science committee that developed the initial 500 000 pounds scientific programme of the institute; peer-reviewing over 100 workshop proposals and hosting over 30. She also chaired the first recruitment wave of the institute hiring 13 data scientists as a multi-university recruitment drive. Sofia was a member of the Royal Society and British Academy Data Governance Working Group, and the Royal Society working group on machine learning. Most recently she was one of 3 commissioners on a law society commission on the usage of algorithms in the justice system.

……………………………………………………………………………………………………

Thursday 24 October 2024, 2-3pm - Victor Chernozhukov (MIT)

This event will take place in CLM 4.02.

Title: Long Story Short: Omitted Variable Bias in Causal Machine Learning.

Abstract: We develop a general theory of omitted variable bias for a wide range of common causal parameters, including (but not limited to) averages of potential outcomes, average treatment effects, average causal derivatives, and policy effects from covariate shifts. Our theory applies to nonparametric models, while naturally allowing for (semi-)parametric restrictions (such as partial linearity) when such assumptions are made. We show how simple plausibility judgments on the maximum explanatory power of omitted variables are sufficient to bound the magnitude of the bias, thus facilitating sensitivity analysis in otherwise complex, nonlinear models. Finally, we provide flexible and efficient statistical inference methods for the bounds, which can leverage modern machine learning algorithms for estimation. These results allow empirical researchers to perform sensitivity analyses in a flexible class of machine-learned causal models using very simple, and interpretable, tools. We demonstrate the utility of our approach with two empirical examples.

Biography:

- Statistician and Economist, with research work focusing on causal inference with high-dimensional data, applications of machine learning methods, counterfactual and policy analysis, distribution and quantile methods, shape restrictions, partial identification, and extreme value theory.
- The International Ford Professor at the Department of Economics and Center for Statistics and Data Science at the MIT, and an International Fellow at CEMMAP, University College London.
- Joined MIT in 2000, after completing Ph.D. in Economics at Stanford University in 2000 and M.S. in Statistics from the UIUC in 1997.
- Also worked as a Senior Principal Scientist for the Core Artificial Intelligence group at Amazon.com for several years, while on academic leave.
- Co-Editor of the Econometrics Journal and an Action Editor of the Journal of Machine Learning Research.
- Elected Fellow of the American Academy of Arts and Sciences, Econometric Society, and Institute of Mathematical Statistics.
- Inaugural Moderator for the Economics section of ArXiv.org launched in 2017.
- Co-author of the new Interdisciplinary Ph.D. program in Statistics at MIT — by the MIT's Institute for Data, Systems, and Society. Co-author of the new B.S. degree 6-14 in Computer Science, Economics, and Data Science at MIT

……

Friday 25 October 2024, 2-3pm - Dennis Lin (Purdue University)

This event will take place in COL.1.06.

Title: AI, BI & SI—Artificial, Biological and Statistical Intelligences

Abstract: Artificial Intelligence (AI) is clearly one of the hottest subjects these days. Basically, AI employs a huge number of inputs (training data), super-efficient computer power/memory, and smart algorithms to perform its intelligence. In contrast, Biological Intelligence (BI) is a natural intelligence that requires very little or even no input. This talk will first discuss the fundamental issue of input (training data) for AI. After all, not-so-informative inputs (even if they are huge) will result in a not-so-intelligent AI. Specifically, three issues will be discussed: (1) input bias, (2) data right vs. right data, and (3) sample vs. population. Finally, the importance of Statistical Intelligence (SI) will be introduced. SI is somehow in between AI and BI. It employs important sample data, solid theoretically proven statistical inference/models, and natural intelligence. In my view, AI will become more and more powerful in many senses, but it will never replace BI. After all, it is said that "The truth is stranger than fiction, because fiction must make sense." The ultimate goal of this study is to find out "how can humans use AI, BI, and SI together to do things better."

Biography: Dr. Dennis K. J. Lin is a University Distinguished Professor and Head of the Statistics Department at Purdue University. His research interests are quality assurance, industrial statistics, data mining, and data science. He has published near 250 SCI/SSCI papers in a wide variety of journals. He currently serves or has served as associate editor for more than 10 professional journals and was co-editor for Applied Stochastic Models for Business and Industry. Dr. Lin is an elected fellow of ASA, IMS, RSS and ASQ, an elected member of ISI, and a lifetime member of ICSA. He is an honorary chair professor for various universities, including Renmin University of China (as a Chang-Jiang Scholar), Fudan University, and National Chengchi University (Taiwan). His recent awards include the 2004 Faculty Scholar Medal Award (Penn State), the Youden Address (ASQ, 2010), the Shewell Award (ASQ, 2010), the Don Owen Award (ASA, 2011), the Loutit Address (SSC, 2011), the Hunter Award (ASQ, 2014), the Shewhart Medal (2015), and the SPES Award at the Joint Statistical Meeting (2018). He will be the 2020 Deming Lecturer at JSM at Philadelphia.

……………………………………………………………………………………

<u>Friday 15 November 2024, 2-3pm - Sahra Ghalebikesabi (Google DeepMind)</u>

This event will take place in the Leverhulme Library, COL.6.15.

Title: Operationalizing Privacy-Conscious Language Model-based Assistants

Abstract: Advanced AI assistants combine frontier LLMs and tool access to autonomously perform complex tasks on behalf of users. While the helpfulness of such assistants can increase dramatically with access to user information including emails and documents, this raises privacy concerns about assistants sharing inappropriate information with third parties without user supervision. To steer information-sharing assistants to behave in accordance with privacy expectations, we propose to operationalize contextual integrity (CI), a framework that equates privacy with the appropriate flow of information in a given context. In particular, we design and evaluate a number of strategies to steer assistants' information-sharing actions to be CI compliant. Our evaluation is based on a novel form filling benchmark composed of synthetic data and human annotations, and it reveals that prompting frontier LLMs to perform CI-based reasoning yields strong results. Paper available at http://arxiv.org/abs/2408.02373.

Biography: Sahra Ghalebikesabi is a Research Scientist at Google DeepMind. Her current research interests revolve around user privacy in LLM-based agents. Before she started at GDM, she was a PhD student in the OxCSML group at the University of Oxford, supervised by Chris Holmes. Her thesis on tackling differential privacy and model misspecification within generative modelling was funded by Microsoft Research PhD fellowship, ESPRC and Novartis.

……………………………………………………………………………………

Friday 22 November 2024, 2-3pm - Francesco Quinzan (University of Oxford)

This event will take place in COL.1.06.

Title: AI Safety and Causal Inference: Challenges and Opportunities

Abstract: Recent successes of AI and Machine Learning have ignited a fast transfer of technology from research into products and government services. This phenomenon has created a range of problems, which can be broadly attributed to the interaction between technology and society. Examples of these problems are bias and unfairness, lack of robustness, and lack of transparency. In this talk, I will discuss some of the main challenges in Trustworthy AI, focusing on various applications, including data-driven health care and RL. I will argue that it is possible to design AI systems that are robust and capable of generalizing effectively, by uncovering the causal mechanisms of the underlying data generating process. I will also discuss how state-of-the-art generative models can be used on top of these techniques, to further enhance generalization performance. I will illustrate recent advancements in this field, and discuss possible future directions.

Biography: Francesco is an associate researcher at the CS Department at the University of Oxford, hosted by Marta Kwiatkowska. He is also an ELSA Research Associate. Previously, Francesco was a Postdoc at the Division of Decision and Control Systems at KTH, where he worked with Stefan Bauer and Cristian Rojas. He obtained his Ph.D. in Computer Science from the Hasso Plattner Institute in Germany. Francesco visited various institutes and research groups, including the Max Plank Institute for Intelligent Systems, where he was hosted by Bernhard Schölkopf, and the Learning & Adaptive Systems Group at ETH. Francesco studied mathematics at the University of Roma Tre, where he graduated with honours.

……………………………………………………………………………………

Thursday 28 November 2024, 2-3pm - Matteo Barigozzi (Università di Bologna)

This event will take place in COL.1.06.

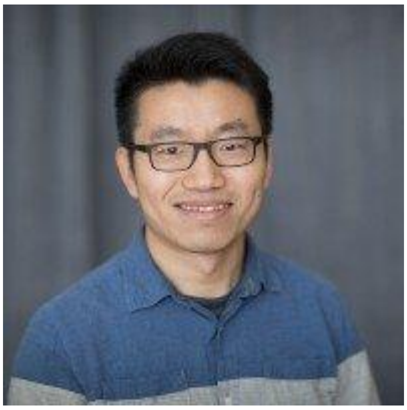
Title: Tail-robust factor modelling of vector and tensor time series in high dimensions

Abstract: We study the problem of factor modelling vector- and tensor-valued time series in the presence of heavy tails in the data, which produce anomalous observations with non-negligible probability. For this, we propose to combine a two-step procedure with data truncation, which is easy to implement and does not require iteratively searching for a numerical solution. Departing away from the light-tail assumptions often adopted in the time series factor modelling literature, we derive the theoretical properties of the proposed estimators while only assuming the existence of the $(2 + 2\eps)$-th moment for some $\eps \in (0, 1)$, fully characterising the effect of heavy tails on the rates of estimation as well as the level of truncation. Numerical experiments on simulated datasets demonstrate the good performance of the proposed estimator, which is further supported by applications to two macroeconomic datasets.

Biography: Matteo Barigozzi is full professor of Political Economy and Econometrics at the department of Economics of the University of Bologna. Before he was Associate professor of Statistics at LSE and post-doc at ECARES (Université libre de Bruxelles). He has a PhD in Economics from Sant'Anna School of Advanced Studies in Pisa and an MSc in Physics from the University of Milano. His current research is on the theory and applications of high-dimensional time series analysis and in particular on large dynamic factor models.

……………………………………………………………………………………………

Friday 29 November 2024, 2-3pm - Tony Qin (Eva AI)



This event will take place in COL.1.06.

Title: Reinforcement Learning & Agents for Marketplace Operations

Abstract: In this talk, we will first review the core developments in reinforcement learning (RL) for the dispatch operations in ride share marketplaces and draw connection to approximate DP through the lens of system value decomposition in the multi-agent setting. We will show how a two-sided view of the graph-based equilibrium metrics (GEM) offers a fundamental way for evaluating marketplace changes. To close with a forward look, we will also describe the emerging problem of using LLM-based agents to automate customer-facing operations for small businesses.

Biography: Tony Qin is Co-founder and Chief Scientist of Eva AI (foreva.ai), building voice AI agentic systems for business operations. Previously, he was Principal Scientist at Lyft Rideshare Labs and Director of the Decision Intelligence group at DiDi AI Labs, spearheaded the development of reinforcement learning (RL) for rideshare marketplace optimization. Tony received his Ph.D. in Operations Research from Columbia University.

He is Associate Editor of the ACM Journal on Autonomous Transportation Systems. He has served as Area Chair/Senior PC of KDD, AAAI, and ECML-PKDD, and a referee of top journals. He is an INFORMS Senior Member, a Franz Edelman Award Finalist and Laureate in 2023 and received the INFORMS Daniel H. Wagner Prize for Excellence in Operations Research Practice in 2019.

....

Friday 6 December 2024, 2 - 2.30pm - Florian Kalinke (Karlsruhe Institute of Technology)

This event will take place in COL.1.06.

Title: Nyström Kernel Stein Discrepancy

Abstract: Testing for goodness-of-fit (GoF) is a classical problem in statistics: Given a known target measure and a set of samples, one wants to test if the samples fit the target. The recent introduction of kernel Stein discrepancies (KSDs) enables principled GoF tests that benefit from the flexibility of reproducing kernel Hilbert spaces and are agnostic to the normalizing constant of the target density, rendering their use especially desirable in Bayesian settings, where the normalizing constant is frequently difficult or even impossible to obtain. However, the runtime of the known U- and V-statistic-based KSD estimators scales quadratically with the number of samples, prohibiting their application in large-scale settings. In this talk, after a quick introduction to KSDs and sub-Gaussianity in Hilbert spaces, I will present the construction of a Nyström-based acceleration and its consistency guarantees, where a sub-Gaussian assumption seems to be the key ingredient. Further, I will demonstrate the applicability of the new algorithm on various GoF benchmarks.

Biography: Florian Kalinke is a fourth-year Ph.D. student in computer science at the Karlsruhe Institute of Technology (KIT) at the "Institute for Program Structures and Data Organization," advised by Klemens Böhm. His research focuses on processing streaming data, online change point detection, the estimation of independence and goodness-of-fit measures, and kernel techniques, with an emphasis on deriving efficient algorithms and analyzing their performance and runtime trade-offs.

....

Friday 6 December 2024, 2.30-3pm - Bharath Sriperumbudur (The Pennsylvania State University)

This event will take place in COL.1.06.

Title: Minimax Optimal Goodness-of-Fit Testing with Kernel Stein Discrepancy

Abstract: We explore the minimax optimality of goodness-of-fit tests on general domains using the kernelized Stein discrepancy (KSD). The KSD framework offers a flexible approach for goodness-of-fit testing, avoiding strong distributional assumptions, accommodating diverse data structures beyond Euclidean spaces, and relying only on partial knowledge of the reference distribution, while maintaining computational efficiency. Although KSD is a powerful framework for goodness-of-fit testing, only the consistency of the corresponding tests has been established so far, and their statistical optimality remains largely unexplored. In this work, we establish a general framework and an operator-theoretic representation of the KSD, encompassing many existing KSD tests in the literature, which vary depending on the domain. Building on this representation, we propose a modified discrepancy by applying the concept of spectral regularization to the KSD framework. We establish the minimax optimality of the proposed regularized test for a wide range of the smoothness parameter θ under a specific alternative space, defined over general domains, using the χ^2-divergence as the separation metric. In contrast, we demonstrate that the unregularized KSD test fails to achieve the minimax separation rate for the considered alternative space. Additionally, we introduce an adaptive test capable of achieving minimax optimality up to a logarithmic factor by adapting to unknown parameters. Through numerical experiments, we illustrate the superior performance of our proposed tests across various domains compared to their unregularized counterparts.

Joint work with Omar Hagrass (Princeton University) and Krishnakumar Balasubramanian (University of California, Davis)

Biography: Bharath Sriperumbudur is currently a Professor of Statistics at Pennsylvania State University. He held a postdoctoral stint at the Gatsby Computational Neuroscience Unit, at University College London, and was a Research Fellow in the Statistical Laboratory, at the University of Cambridge. He received his Ph. D. in Electrical Engineering from the University of California, San Diego. He is the recipient of the prestigious NSF CAREER Award. He is currently serving as an Action Editor for the Journal of Machine Learning Research and has served as an area chair for many machine learning conferences such as NeurIPS, COLT, ALT, ICML, and AISTATS. His current research interests are in statistical learning theory, non-parametric statistics, RKHS theory and methods, topological data analysis, optimal transport, and gradient flows.

………………………………………………………………………………………

Thursday 12 December 2024, 2-3pm - Kangjie Zhou (Columbia University)

This event will take place in COL.1.06.

Title: Dynamic Factor Analysis of High-dimensional Recurrent Events.

Abstract: Recurrent event time data arise in many studies, including biomedicine, public health, marketing, and social media analysis. High-dimensional recurrent event data involving large numbers of event types and observations become prevalent with the advances in information technology. This paper proposes a semiparametric dynamic factor model for the dimension reduction and prediction of high-dimensional recurrent event data. The proposed model imposes a low-dimensional structure on the mean intensity functions of the event types while allowing for dependencies. A nearly rate-optimal smoothing-based estimator is proposed. An information criterion that consistently selects the number of factors is also developed. Simulation studies demonstrate the effectiveness of these inference tools. The proposed method is applied to grocery shopping data, for which an interpretable factor structure is obtained. Based on joint work with Fangyi Chen, Yunxiao Chen and Zhiliang Ying.

Biography: Kangjie Zhou is a Founder’s postdoctoral researcher in the Department of Statistics at Columbia University. He obtained PhD in statistics from Stanford University in May 2024, where he was advised by Andrea Montanari. His research interests span theoretical statistics, machine learning, and probability.