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

Autumn Term 2023

Friday 13 October 2023, 2-3pm - Bharath Sriperumbudur (Pennsylvania State University)

This event will take place in the Data Science Institute (COL.1.06).

Title: Gromov-Wasserstein Distances: Entropic Regularization, Duality, and Sample Complexity

Abstract: The Gromov-Wasserstein (GW) distance, rooted in optimal transport (OT) theory, quantifies dissimilarity between metric measure spaces and provides a framework for aligning heterogeneous datasets. While computational aspects of the GW problem have been widely studied, a duality theory and fundamental statistical questions concerning empirical convergence rates remained obscure. In this talk, I present our recent work that closes these gaps for the quadratic GW distance over Euclidean spaces of different dimensions d_x and d_y. We treat both the standard and the entropically regularized GW distance, and derive dual forms that represent them in terms of the well-understood OT and entropic OT (EOT) problems, respectively. This enables employing proof techniques from statistical OT based on regularity analysis of dual potentials and empirical process theory, using which we establish the first GW empirical convergence rates. The derived two-sample rates are $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ (up to a log factor when $\min\{d_x,d_y\}=4$) for standard GW and $n^{-1/2}$ for EGW, which matches the corresponding rates for standard and entropic OT. The parametric rate for EGW is evidently optimal, while for standard GW we provide matching lower bounds, which establish sharpness of the derived rates. We also study stability of EGW in the entropic regularization parameter and prove approximation and continuity results for the cost and optimal couplings. Our results serve as a first step towards a comprehensive statistical theory as well as computational advancements for GW distances, based on the discovered dual formulations.
Joint work with Zhengxin Zhang (Cornell University, Applied Math), Ziv Goldfeld (Cornell University, ECE), and Youssef Mroueh (IBM Research, NYC)

Biography: Bharath Sriperumbudur is currently an Associate Professor of Statistics at Pennsylvania State University. He held a postdoctoral stint at Gatsby

Computational Neuroscience Unit, University College London, and was a Research Fellow in the Statistical Laboratory, University of Cambridge. He received his Ph. D. in Electrical Engineering from 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.

Take a look at Bharath's slides (PDF).

……

Friday 20 October 2023, 2-3pm - Florian Kalinke (Karlsruhe Institute of Technology)

This event will take place in B.07, Sir Arthur Lewis Building (SAL.B.07).

Title: Nyström M-Hilbert-Schmidt Independence Criterion

Abstract: Measuring the joint independence of random variables is a central problem in data science, with Hilbert-Schmidt independence criterion (HSIC) being one of the most popular measures in the area. While closed-form estimators for HSIC exist, they scale quadratically with the number of samples, which proves to be prohibitive for large-scale applications. Accordingly, speed-ups for the estimation of HSIC exist, but they are limited to two random variables. In this talk, I will present a Nyström-based accelerated estimation of HSIC that allows more than two random variables, accompanied by consistency and mini-max optimality guarantees. I will demonstrate the applicability of the approach for the dependency testing of media annotations and causal discovery.

Biography: Florian Kalinke is a third-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, estimation of independence measures, and kernel techniques, with a particular focus on deriving efficient algorithms and the analysis of their performance and runtime trade-offs, and applying these techniques in energy management.

Take a look at Florian's slides (PDF).

……..

Friday 27 October 2023, 2-3pm - Claire Monteleoni (INRIA Paris & University of Colorado Boulder)

This event will take place in the Data Science Institute (COL.1.06).

Title: Machine Learning Research for Climate Change and Environmental Sustainability

Abstract: Despite the scientific consensus on climate change, drastic uncertainties remain. Crucial questions about regional climate trends, changes in extreme events, such as heat waves and mega-storms, and understanding how climate varied in the distant past, must be answered in order to improve predictions, assess impacts and vulnerability, and inform mitigation and sustainable adaptation strategies. Machine learning can help answer such questions and shed light on climate change. I will give an overview of our climate informatics research, focusing on challenges in learning from spatiotemporal data, along with semi- and unsupervised deep learning approaches to studying rare and extreme events, and precipitation and temperature downscaling

Biography: Claire Monteleoni is a Choose France Chair in AI and a Research Director at INRIA Paris, a Professor in the Department of Computer Science at the University of Colorado Boulder, and the founding Editor in Chief of Environmental Data Science, a Cambridge University Press journal, launched in December 2020. She joined INRIA in 2023 and has previously held positions at University of Paris-Saclay, CNRS, George Washington University, and Columbia University. She completed her PhD and Masters in Computer Science at MIT and was a postdoc at UC San Diego. She holds a Bachelor's in Earth and Planetary Sciences from Harvard. Her research on machine learning for the study of climate change helped launch the interdisciplinary field of Climate Informatics. She co-founded the International Conference on Climate Informatics, which turned 12 years old in 2023, and has attracted climate scientists and data scientists from over 20 countries and 30 U.S. states. She gave an invited tutorial: Climate Change: Challenges for Machine Learning, at NeurIPS 2014. She currently serves on the NSF Advisory Committee for Environmental Research and Education.

Take a look at Claire's slides (PDF).

……

Friday 10 November 2023, 2-3pm - Emanuele Giorgi (Lancaster University)

This event will take place in the Data Science Institute (COL.1.06).

Title: Model-based geostatistical inference with low prevalence data: a case study on lymphatic filariasis

Abstract: Model-based geostatistics is a branch of spatial statistics that is used to draw inferences on a sptially continuous surface using data collected at a discrete set of locations. In this presentation, I will first give an overview of the concepts underpinning MBG and its application to public health problems where the goal is to predict disease risk within an geographical area of interests. I will then introduce the problem of analysing low-prevalence disease data were, due to the large number of zero reported cases, the estimation of MBG models can be quite challenging. This problem often arises in the context of diseases approaching elimination and where MBG has been used to assess the elimination status of areal units. Through a case study on lymphatic filariasis, a mosquito-borne parasitic disease that causes chronic and disabling swelling of the lymphatic system, we will present a spatio-temporal MBG model that uses historical data to draw inferences on the prevalence surface in a post-elimination setting. We will also illustrate how this approach can be used to design a surveillance system to quantify the likelihood of the resurgence of the disease in a post-elimination phase.

Biography: Emanuele Giorgi is a Senior Lecturer in Biostatistics at Lancaster University. He leads the Centre for Health for Health Informatics Computing and Statistics (CHICAS) which is a designated WHO Collaborating Centre on Geostatistical Methods for Neglected Tropical Disease Research.His primary research interests are in the development of spatial and spatio-temporal statistical methodologies and their implementation into open-source statistical software. His application domain is global health, with a specific focus on tropical diseases in low and middle-income countries.

Take a look at Emanuele's slides (PDF).

……

Monday 20 November 2023, 2-3pm - Zhigang Yao (National University of Singapore

This event will take place in the Data Science Institute (COL.1.06).

Title: Manifold Fitting -An Invitation to Statistics

Abstract: This manifold fitting problem can go back to H. Whitney's work in the early 1930s (Whitney (1992)), and finally has been answered in recent years by C. Fefferman's works (Fefferman, 2006, 2005). The solution to the Whitney extension problem leads to new insights for data interpolation and inspires the formulation of the Geometric Whitney Problems (Fefferman et al. (2020, 2021a)): Assume that we are given a set $Y \subset \mathbb{R}^D$. When can we construct a smooth d-dimensional submanifold $\widehat{M} \subset \mathbb{R}^D$ to approximate Y, and how well can $\widehat{M}$ estimate Y in terms of distance and smoothness? To address these problems, various mathematical approaches have been proposed (see Fefferman et al. (2016, 2018, 2021b)). However, many of these methods rely on restrictive assumptions, making extending them to efficient and workable algorithms challenging. As the manifold hypothesis (non-Euclidean structure exploration) continues to be a foundational element in statistics, the manifold fitting Problem, merits further exploration and discussion within the modern statistical community. The talk will be partially based on recent works of Yao and Xia (2019) and Yao, Su, Li and Yau (2022) and Yao, Su, Yau (2023)

Biography: Zhigang Yao is a tenured associate professor in the Department of Statistics and Data Science at the National University of Singapore. He has been a visiting faculty at the Center for Mathematical Sciences and Applications at Harvard University since 2022. He holds a visiting professorship at YMSC at Tsinghua University. His primary research interests lie in statistical inference for complex data. In recent years, his focus has shifted towards Non-Euclidean Statistics and low-dimensional manifold learning.Yao is committed to promoting the new field of interaction between geometry and statistics. In recent years, Yao and his collaborators have proposed methods and theories that redefine traditional PCA on Riemannian manifolds including principal flows/sub-manifolds and principal boundaries, as well as new manifold learning theories. These methods aim to address deficiencies in traditional statistical methods and theories by taking into account the geometry of the data.

……..

Friday 24 November 2023, 2-3pm - Daniela Witten (University of Washington)

This event will take place in the Data Science Institute (COL.1.06).

Title: Data thinning and its applications

Abstract: We propose data thinning, a new approach for splitting an observation from a known distributional family with unknown parameter(s) into two or more independent parts that sum to yield the original observation, and that follow the same distribution as the original observation, up to a (known) scaling of a parameter. This proposal is very general, and can be applied to a broad class of distributions within the natural exponential family, including the Gaussian, Poisson, negative binomial, Gamma, and binomial distributions, among others. Furthermore, we generalize data thinning to enable splitting an observation into two or more parts that can be combined to yield the original observation using an operation other than addition; this enables the application of data thinning far beyond the natural exponential family. Data thinning has a number of applications to model selection, evaluation, and inference. For instance, cross-validation via data thinning provides an attractive alternative to the "usual" approach of cross-validation via sample splitting, especially in unsupervised settings in which the latter is not applicable. We will present an application of data thinning to single-cell RNA-sequencing data, in a setting where sample splitting is not applicable. This is joint work with Anna Neufeld (Fred Hutch), Ameer Dharamshi (University of Washington), Lucy Gao (University of British Columbia), and Jacob Bien (University of Southern California)

……

Friday 1 December 2023, 2-3pm - Giulia Martini (United Nations World Food Programme)

This event will take place in the Data Science Institute (COL.1.06).

Title: Food security monitoring, from real time data to machine learning

Abstract: Hunger Monitoring harnesses mobile technology, artificial intelligence and data analytics to establish remote, near real-time food security monitoring systems across countries to enable WFP, governments, partners and the broad humanitarian

community to monitor the food security situation daily, capture problems in real-time in the event of a crisis and provide the necessary information for early action and mitigation.
By bringing food security analysts with expertise in data science, engineering and predictive modelling together, we collect daily data on hunger and its drivers and observe how situations may shift from one day to the next to inform effective response. To ensure that near real-time data is available as a global public good, we developed the HungerMapLIVE, WFP's global hunger monitoring system so that anyone, anywhere can track hunger as it is now in over 90 countries. Survey data is used to fit machine learning models based on the XGBoost algorithm to make daily estimates of the food insecurity levels with first administrative area resolution.

Biography: Giulia is a data scientist who currently works at the United Nations World Food Program (WFP) in the Hunger Monitoring Unit where she leverages her expertise to develop and implement machine learning models. These models are used to estimate food security indicators in regions where primary data is unavailable and to provide longer-term forecasting. Before joining WFP in 2020, she worked as a researcher at the Netherlands Organisation for Applied Scientific Research (TNO) and received a Master's degree in civil engineering from Delft University of Technology. Giulia's primary objective is to make technology accessible and user-friendly to support and improve operations in the humanitarian sector.

…….

Friday 8 December 2023, 2-3pm - Maria-Pia Victoria Feser (University of Geneva)

This event will take place in B.07, Sir Arthur Lewis Building (SAL.B.07).

Title: Fast M-Estimation of Generalized Linear Latent Variable Models in High Dimensions

Abstract: Dimension reduction for high dimensional data is an important and challenging task, relevant to both machine learning and statistical applications. Generalized Linear Latent Variable Models (GLLVMs) provide a probabilistic alternative to matrix factorization when the data are of mixed types, whether discrete, continuous, or a mixture of both. They achieve the reduction of dimensionality by mapping the correlated multivariate data to so-called latent variables, defined in a lower-dimensional space. The benefit of GLLVMs is twofold:

the latent variables can be estimated and used as features to be embedded in another model, and the model parameters themselves are interpretable and provide meaningful indications on the very structure of the data. Moreover, GLLVM can naturally be extended to dynamic processes such as those used to model longitudinal data. However, with a likelihood based approach, GLLVM's estimation represents a tremendous challenge for even moderately large dimensions, essentially due to the multiple integrals involved in the likelihood function. Numerous methods based on approximations of this latter have been proposed: Laplace approximation, adaptive quadrature, or, recently, extended variational approximation. For GLLVMs, however, these methods do not scale well to high dimensions, and they may also introduce a large bias in the estimates. In this paper, we consider an alternative route, which consists in proposing an alternative estimator, based on drastically simplified estimating equations, complemented with a numerically efficient bias reduction methods in order to recover a consistent estimator for the GLLVM parameters. The resulting estimator is an M-estimator, which has a negligible efficiency loss compared to the (exact) MLE. For larger data sets, the proposed M-estimator, whose computational burden is linear in npq, remains applicable when the state-of-the-art method fails to converge. To compute the M-estimator, we propose to use a stochastic approximation algorithm.

Biography: Graduated from the University of Geneva (Ph. D. in econometrics and statistics) in 1993, Maria-Pia Victoria-Feser has held several positions in different institutions or Departments. She was appointed as lecturer in statistics at the London School of Economics (1993-1996), as assistant and associate professor in statistics (part time) at the Faculty of psychology and educational sciences at the University of Geneva (1997-2005), financed by a Swiss National Science Found grant, full professor in statistics at the University of Geneva since 2001. She now holds, since 2023, a full professor position at the department of statistics of the University of Bologna. She has also acted for the foundation and as founding dean (2013-2017) of the Geneva School of Economics and Management (GSEM) of the University of Geneva, and as founding director of the Research Center for Statistics of the University of Geneva (created in 2011).

Maria-Pia Victoria-Feser's research interests are in fundamental statistics (robust statistics, model selection and simulation based inference in high dimensions for complex models) with applications in economics (welfare economics, extremes), psychology and social sciences (generalized linear latent variable models, media

analytics), and engineering (time series for geo-localization). She has published in leading journals of statistics as well as top journals in related fields. Throughout her career, she has supervised several PhD students, three of which currently hold professorial positions and one a senior researcher position, in leading academic institutions.

Take a look at Maria-Pia's slides (PDF).

……

Winter Term 2024

Friday 2 February 2024, 2-3pm - Tom Everitt (Google Deepmind)

This event has been postponed - details to follow.

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.

……

Monday 4 March 2024, 2-3pm - Claire Donnat (University of Chicago)

This event will take place in the Data Science Institute (COL.1.06).

Title: Sparse topic modeling via spectral decomposition and thresholding

Abstract: By modeling documents as mixtures of topics, Topic Modeling allows the discovery of latent thematic structures within large text corpora, and has played an important role in natural language processing over the past decades. Beyond text data, topic modeling has proven itself central to the analysis of microbiome data,

population genetics, or, more recently, single-cell spatial transcriptomics. Given the model's extensive use, the development of estimators — particularly those capable of leveraging known structure in the data— presents a compelling challenge.

In this talk, we focus more specifically on the probabilistic Latent Semantic Indexing model, which assumes that the expectation of the corpus matrix is low-rank and can be written as the product of a topic-word matrix and a word-document matrix. Although various estimators of the topic matrix have recently been proposed, their error bounds highlight a number of data regimes in which the error can grow substantially— particularly in the case where the size of the dictionary p is large.

In this talk, we propose studying the estimation of the topic-word matrix under the assumption that the ordered entries of its columns rapidly decay to zero. This sparsity assumption is motivated by the empirical observation that the word frequencies in a text often adhere to Zipf's law. We introduce a new spectral procedure for estimating the topic-word matrix that thresholds words based on their corpus frequencies, and show that its ℓ1-error rate under our sparsity assumption depends on the vocabulary size p only via a logarithmic term. Our error bound is valid for all parameter regimes and in particular for the setting where p is extremely large; Our procedure also empirically performs well relative to well-established methods when applied to a large corpus of research paper abstracts, as well as the analysis of single-cell and microbiome data where the same statistical model is relevant but the parameter regimes are vastly different.

Bio: Claire Donnat is an Assistant Professor in the Department of Statistics at the University of Chicago. She obtained her PhD in Statistics from Stanford in 2020, where she worked with Prof. Susan Holmes. Her work focuses on statistical estimation in high dimensions with structure, typically encoded by a graph and motivated by applications in biology.

……

Friday 15 March 2024, 2-3pm - Henry Reeve (University of Bristol)

This event will take place in the Data Science Institute (COL.1.06).

Title: Non-stationary label shift and adaptive estimation

Abstract: We shall consider a non-stationary setting in which the learner has access to both a labelled and an unlabelled sample. Our standing assumption for this classification problem will be label-shift: Whilst the marginal probabilities of the classes may vary over time, the class-conditional distributions are stationary. Our objective is to combine information from a fixed labelled sample drawn from the class-conditional distributions with an evolving unlabelled sample drawn from a non-stationary distribution. We highlight the advantage of a methodology based on adaptive estimation for non-stationary problems, in contrast to widely used techniques based on hedging. Our methodology draws upon a refinement of the Massart-Dvoretzky–Kiefer–Wolfowitz inequality which we believe to be of independent interest.

Bio: Henry Reeve is a Lecturer in Statistical Science at the University of Bristol. His research focuses on Statistics and Machine Learning, including transfer learning, subgroup discovery, compressive learning, classification with label noise and multi-armed bandits. Henry received the Steve Furber Medal for his PhD in Computer Science from the University of Manchester and completed his postdoctoral research with Ata Kaban at the University of Birmingham.

……

Spring Term 2024

Friday 3 May 2024, 2-3pm - Yuhao Wang (Tsinghua University)

This event will take place in the Data Science Institute (COL.1.06).

Title: Residual Permutation Test for High-Dimensional Regression Coefficient Testing

Abstract: We consider the problem of testing whether a single coefficient is equal to zero in fixed-design linear models under a moderately high-dimensional regime, where the dimension of covariates p is allowed to be in the same order of magnitude as sample size n. In this regime, to achieve finite-population validity, existing methods usually require strong distributional assumptions on the noise vector (such as Gaussian or rotationally invariant), which limits their applications in practice. In this paper, we propose a new method, called residual permutation test (RPT), which is constructed by projecting the regression residuals onto the space orthogonal to

the union of the column spaces of the original and permuted design matrices. RPT can be proved to achieve finite-population size validity under fixed design with just exchangeable noises, whenever p < n/2. Moreover, RPT is shown to be asymptotically powerful for heavy tailed noises with bounded (1+t)-th order moment when the true coefficient is at least of order n^{−t/(1+t)} for t \in [0,1]. We further proved that this signal size requirement is essentially rate-optimal in the minimax sense. Numerical studies confirm that RPT performs well in a wide range of simulation settings with normal and heavy-tailed noise distributions. This is based on joint works with Tengyao Wang and Kaiyue Wen.

Bio: Yuhao Wang is an assistant professor in the Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University. Before joining Tsinghua, Yuhao was a postdoctoral research associate at the Statistical Laboratory, which is part of the Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge. Yuhao received his Ph.D. from the Department of Electrical Engineering and Computer Science at Massachusetts Institute of Technology. Prior to Ph.D., Yuhao got his bachelor from the Department of Automation in Tsinghua University.

Take a look at Yuhao's slides (PDF).

……

Friday 10 May 2024, 2-3pm - Sinead Williamson (University of Texas)

This event will take place in the Data Science Institute (COL.1.06).

Title: Posterior Uncertainty Quantification in Neural Networks using Data Augmentation

Abstract: We approach the problem of uncertainty quantification in deep learning through a predictive framework, which captures uncertainty in model parameters by specifying our assumptions about the predictive distribution of unseen future data. Under this view, we show that deep ensembling (Lakshminarayanan et al., 2017) is a fundamentally mis-specified model class, since it assumes that future data are supported on existing observations only -- a situation rarely encountered in practice. To address this limitation, we propose MixupMP, a method that constructs a more realistic predictive distribution using popular data augmentation techniques. MixupMP operates as a drop-in replacement for deep ensembles, where each

ensemble member is trained on a random simulation from this predictive distribution. Grounded in the recently-proposed framework of Martingale posteriors (Fong et al., 2023), MixupMP returns samples from an implicitly defined Bayesian posterior. Our empirical analysis showcases that MixupMP achieves superior predictive performance and uncertainty quantification on various image classification datasets, when compared with existing Bayesian and non-Bayesian approaches.

Joint work with Luhuan Wu

Bio: Sinead Williamson is a researcher in Apple's Machine Learning Research team. Prior to joining Apple, she was an Associate Professor in the Department of Statistics and Data Sciences, at The University of Texas at Austin. Sinead's main research focus is the development of Bayesian methods for machine learning applications. In particular, she is interested in constructing distributions over correlated measures and complex structures, in order to model structured data sets or data with spatio-temporal dependence. Examples range from temporally evolving social networks, to complex image or language data.

……..

Monday 2 September 2024, 2-3pm - Qiyang Han (Rugters University)

This event will take place in the Data Science Institute (COL.1.06).

Title: Entrywise Dynamics and Universality of General First Order Methods

Abstract: General first order methods (GFOMs), including many variants of gradient descent and approximate message passing algorithms, constitute a broad class of iterative algorithms widely applied in modern statistical learning problems. Some GFOMs also serve as constructive proof devices, iteratively characterizing the empirical distributions of statistical estimators in the asymptotic regime of large system limits for any fixed number of iterations.

This paper develops a non-asymptotic, entrywise characterization of the dynamics for a general class of GFOMs. Our characterizations capture the precise stochastic behavior of each coordinate of the GFOM iterates, and more importantly, hold universally across a broad class of heterogeneous random matrix models. As a

corollary, we provide the first non-asymptotic description of the empirical distributions of the GFOM iterates beyond Gaussian ensembles.

We demonstrate the utility of our general GFOM theory through two sets of applications. In the first application, we develop a new algorithmic approach to prove universality for general empirical risk minimizers. Specifically, we establish new entrywise universality for a broad class of regularized least squares estimators in the linear model, by controlling the entrywise error relative to a suitably constructed GFOM iterate. This algorithmic proof method also systematically improves averaged universality results for general regularized regression estimators in the linear model, and resolves the universality conjecture for (regularized) maximum likelihood estimators in the logistic regression model. In the second application, we obtain entrywise Gaussian approximations for a general class of gradient descent algorithms. Our approach provides non-asymptotic state evolution for the bias and variance of the algorithm along the iteration path, applicable even to non-convex loss functions.

Bio: Qiyang Han received the Ph.D. degree from the University of Washington, Seattle in 2018. He is currently an Associate Professor of statistics at Rutgers, the State University of New Jersey. He is broadly interested in mathematical statistics and high-dimensional probability, with a particular focus on empirical process theory and its applications to nonparametric and high dimensional statistics. He received the Bernoulli Society New Researcher Award and the David G. Kendall Award in 2023 for his contribution to mathematical statistics.

End of Archive.