Programme

Monday, 10th June

- 8:45am–9:00am **Registration**
- 9:00am–9:10am **Welcome**
- 9:10am–10:00am
  - Henry Reeve: Adaptive classification for non-stationary label-shift and a local Dvoretzky–Kiefer–Wolfowitz–Massart inequality
- **Coffee Break**
- 10:30am–11:20am
  - Kabir Aladin Verchand: Mean estimation beyond missing completely at random
- 11:30am–12:20pm
  - Claire Boyer: Linear prediction with missing inputs: a blessing or a curse?
- **Lunch Break**
- 2pm–2:50pm
  - Angelika Rohde: Nonparametric Bootstrap of High-Dimensional Sample Covariance Matrices
- 3pm–3:50pm
  - Torben Sell: Nonparametric classification with missing data
- **Tea Break**
- 4:30pm–5:20pm
  - Weijie Su: A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules
- 6pm–8pm **Reception and Poster Session**

Tuesday, 10th September

- 9:00am–9:50am
  - Richard Samworth: Optimal convex M-estimation via score matching
- **Coffee Break**

- 10:30am–11:20am
  - Olga Klopp: Denoising over network with application to partially observed epidemics
- 11:30am–12:20pm
  - Ariane Marandon: Selecting informative conformal prediction sets with false coverage rate control
- **Lunch Break**
- 2pm–2:50pm
  - Botond Szabo: On distributed estimation and testing under communication constraints
- 3pm–3:50pm
  - Yi Yu: Local differential privacy: User-level and federated learning
- **Tea Break**
- 4:30pm–5:20pm
  - Dominik Rothenhäusler: Out-of-distribution generalization under random, dense distributional shifts
- 6:30pm **Conference Dinner** (Scarman, by invitation only)

Wednesday, 11th September

- 9:00am–9:50am
  - Niklas Pfister: Extrapolation-Aware Nonparametric Statistical Inference
- 9:50am–10:40am
  - Yuhao Wang: Long-term Causal Inference Under Persistent Confounding via Data Combination
- **Coffee Break**
- 11:00am–11:50am
  - Mona Azadkia: A simple measure of conditional dependence
- **Lunch Break**
- 1:30pm–4pm
  - Open Discussion

Talk abstracts

Kabir Aladin Verchand (University of Cambridge)
Monday 10:30am

Title: Mean estimation beyond missing completely at random

Abstract: We consider mean estimation in the presence of missing data. Moving beyond restrictive missing completely at random (MCAR) assumptions, we first formulate a missing data analogue of Huber's contamination model. We then show that, for mean estimation, the minimax rate of convergence decomposes into a sum of the minimax rate under a heterogeneous, MCAR assumption and a robust error term incurred by departure from MCAR.

We next re-visit the class of contamination distributions and introduce the finer notion of realisability, which is rich enough to capture all missing not at random distributions constructed from a base distribution. We show that estimation can be considerably easier under realisable contaminations. For instance, in the univariate setting, we show that if the base distribution is Gaussian, then our minimax optimal estimators can consistently estimate the mean. Further, under realisable contamination, our estimators enjoy a high breakdown point converging to 1.

Mona Azadkia (London School of Economics)
Wednesday 11am

Title: A simple measure of conditional dependence

Abstract: We propose a coefficient of conditional dependence between two random variables Y and Z given a set of other variables $X_1, \ldots, X_p$, based on an i.i.d. sample. The coefficient has a long list of desirable properties, the most important of which is that under absolutely no distributional assumptions, it converges to a limit in $[0, 1]$, where the limit is 0 if and only if Y and Z are conditionally independent, given $X_1, \ldots, X_p$, and is one if and only if Y is equal to a measurable function of Z given $X_1, \ldots, X_p$. Moreover, it has a natural interpretation as a nonlinear generalization of the familiar partial R^2 statistic for measuring conditional dependence by regression. Using this statistic, we devise a new variable selection algorithm called Feature Ordering by Conditional Independence (FOCI), which is model-free, has no tuning parameters, and is provably consistent under sparsity assumptions. Finally, we discuss recent advances and generalizations of the statistics.

Claire Boyer (Sorbonne Université)
Monday 11:30am

Title: Linear prediction with missing inputs: a blessing or a curse?

Abstract: Two different approaches exist to handle missing values for prediction: either imputation, prior to fitting any predictive algorithms, or dedicated methods able to natively incorporate missing values. For the latter, one can think about pattern by pattern strategies where a predictor can be learned depending on the missing pattern of the test sampe. We will discuss both approaches to make linear models handle missing inputs.

Olga Klopp (ESSEC Business School)
Tuesday 10:30am

Title: Denoising over network with application to partially observed epidemics

Abstract: We introduce a novel approach to predict epidemic spread over networks using total variation (TV) denoising, a signal processing technique. The study proves the consistency of TV denoising with Bernoulli noise, extending existing bounds from Gaussian noise literature. The methodology is further extended to handle incomplete observations, showcasing its effectiveness. We show that application of 1-bit total variation denoiser improves the prediction accuracy of virus spread dynamics on networks.

Ariane Marandon (Turing Institute)
Tuesday 11:30am

Title: Selecting informative conformal prediction sets with false coverage rate control

Abstract: In supervised learning, including regression and classification, conformal methods provide prediction sets for the outcome/label with finite sample coverage for any machine learning predictor. We consider here the case where such prediction sets come after a selection process which requires that the selected prediction sets be "informative" in a well-defined sense. We consider both the classification and regression settings where the analyst may consider as informative only the prediction sets small enough, excluding null values, or obeying other appropriate "monotone" constraints. We develop a unified framework for building such informative conformal prediction sets while controlling the false coverage rate (FCR) on the selected.

Niklas Pfister (University of Copenhagen)
Wednesday 9am

Title: Extrapolation-Aware Nonparametric Statistical Inference

Abstract: Extrapolation occurs in many data analysis applications and can invalidate the resulting conclusions if not taken into account. Formally, extrapolation refers to any type of statistical inference on a conditional function (e.g., a conditional expectation or conditional quantile) evaluated outside of the support of the conditioning variable. While extrapolation is straightforward in parametric models, it becomes challenging in nonparametric models. In this talk, we extend the nonparametric statistical model to explicitly allow for extrapolation and introduce a class of extrapolating assumptions that can be combined with existing inference techniques to draw extrapolation-aware conclusions. The proposed class of extrapolation assumptions stipulate that the conditional function of interest attains its minimal and maximal directional derivative in each direction within the observed support. We illustrate how the framework can be applied to several statistical applications including out-of-support prediction and extrapolation-aware uncertainty quantification.

Henry Reeve (University of Bristol)
Monday 9:10am

Title: Adaptive classification for non-stationary label-shift and a local Dvoretzky–Kiefer–Wolfowitz–Massart inequality

Abstract: We consider a semi-supervised classification problem with non-stationary label-shift in which we observe a labelled data set followed by a sequence of unlabelled covariate vectors in which the marginal probabilities of the class labels may change over time. Our objective is to predict the corresponding class-label for each covariate vector, without observing the ground-truth labels, beyond the initial labelled data set. Previous work has demonstrated the potential of hedging based strategies to perform competitively with the optimal dynamic strategy (Bai et al. 2022). In this work we explore an alternative approach grounded in adaptive estimation. This approach leverages a recent local Dvoretzky–Kiefer–Wolfowitz–Massart inequality which holds uniformly over sub-intervals of the real line with an error rate that adapts to the behaviour of the population distribution function on the interval. We demonstrate the merits of this alternative methodology by establishing a high-probability regret bound on the test error at any given individual test-time, which adapt automatically to the unknown dynamics of the marginal label probabilities. Further more, we give bounds on the average dynamic regret which match the average guarantees of the online gradient descent based method for any given time interval.

Angelika Rohde (Universität Freiburg)
Monday 2pm

Title: Nonparametric Bootstrap of High-Dimensional Sample Covariance Matrices

Abstract: We introduce a new "$(m, mp/n)$ out of (n, p)"-sampling with replacement bootstrap for eigenvalue statistics of high-dimensional sample covariance matrices based on n independent p-dimensional random vectors. In the high-dimensional scenario $p/n \to c \in (0, \infty)$, this fully nonparametric and computationally tractable bootstrap is shown to consistently reproduce the underlying spectral measure if $m/n \to 0$. If $m^2/n \to 0$, it approximates correctly the distribution of linear spectral statistics. The crucial component is a suitably defined representative subpopulation condition which is shown to be verified in a large variety of situations. Our proofs are conducted under minimal moment requirements and incorporate delicate results on non-centered quadratic forms, combinatorial trace moments estimates as well as a conditional bootstrap martingale CLT which may be of independent interest.

Dominik Rothenhäusler (Stanford University)
Tuesday 4:30pm

Title: Out-of-distribution generalization under random, dense distributional shifts

Abstract: Many existing approaches for estimating parameters in settings with distributional shifts operate under an invariance assumption. For example, under covariate shift, it is assumed that p(y—x) remains invariant. We refer to such distribution shifts as sparse, since they may be substantial but affect only a part of the system. In contrast, in various real-world settings, shifts might be dense. More specifically, these shifts may arise through many small and random changes in the population and environment. First, we will discuss empirical evidence for such random dense distributional shifts and explain why commonly used models for distribution shifts—including adversarial approaches—may not be appropriate under these conditions. Then, we will develop tools to infer parameters and make predictions for partially observed, shifted distributions. Finally, we will apply the framework to several real-world datasets and discuss diagnostics to evaluate the fit of the distributional uncertainty model.

Richard Samworth (University of Cambridge)
Tuesday 9am

Title: Optimal convex M-estimation via score matching

Abstract: In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. Our semiparametric approach targets the best decreasing approximation of the derivative of the log-density of the noise distribution. At the population level, this fitting process is a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. The procedure is computationally efficient, and we prove that our procedure attains the minimal asymptotic covariance among all convex M-estimators. As an example of a non-log-concave setting, for Cauchy errors, the optimal convex loss function is Huber-like, and our procedure yields an asymptotic efficiency greater than 0.87 relative to the oracle maximum likelihood estimator of the regression coefficients that uses knowledge of this error distribution; in this sense, we obtain robustness without sacrificing much efficiency.

Torben Sell (University of Edinburgh)
Monday 3pm

Title: Nonparametric classification with missing data

Abstract: We introduce a new nonparametric framework for classification problems in the presence of missing data. The key aspect of our framework is that the regression function decomposes into an anova-type sum of orthogonal functions, of which some (or even many) may be zero. Working under a general missingness setting, which allows features to be missing not at random, our main goal is to derive the minimax rate for the excess risk in this problem. In addition to the decomposition property, the rate depends on parameters that control the tail behaviour of the marginal feature distributions, the smoothness of the regression function and a margin condition. The ambient data dimension does not appear in the minimax rate, which can therefore be faster than in the classical nonparametric setting. We further propose a new method, called the Hard-thresholding Anova Missing data (HAM) classifier, based on a careful combination of a k-nearest neighbour algorithm and a thresholding step. The HAM classifier attains the minimax rate up to polylogarithmic factors and numerical experiments further illustrate its utility.

Weijie Su (University of Pennsylvania)
Monday 4:30pm

Title: A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules

Abstract: Since ChatGPT was introduced in November 2022, embedding (nearly) unnoticeable statistical signals into text generated by large language models (LLMs), also known as watermarking, has been used as a principled approach to provable detection of LLM-generated text from its human-written counterpart. In this talk, we will introduce a general and flexible framework for reasoning about the statistical efficiency of watermarks and designing powerful detection rules. Inspired by the hypothesis testing formulation of watermark detection, our framework starts by selecting a pivotal statistic of the text and a secret key – provided by the LLM to the verifier – to enable controlling the false positive rate (the error of mistakenly detecting human-written text as LLM-generated). Next, this framework allows one to evaluate the power of watermark detection rules by obtaining a closed-form expression of the asymptotic false negative rate (the error of incorrectly classifying LLM-generated text as human-written). Our framework further reduces the problem of determining the optimal detection rule to solving a minimax optimization program. We apply this framework to two representative watermarks – one of which has been internally implemented at OpenAI – and obtain several findings that can be instrumental in guiding the practice of implementing watermarks. In particular, we derive optimal detection rules for these watermarks under our framework. These theoretically derived detection rules are demonstrated to be competitive and sometimes enjoy a higher power than existing detection approaches through numerical experiments.

Botond Szabo (Bocconi University)
Tuesday 2pm

Title: On distributed estimation and testing under communication constraints

Abstract: In recent years, the amount of available information has become so vast in certain fields of applications that it is infeasible or undesirable to carry out all the computations on a single server. This has motivated the design and study of distributed statistical or learning approaches. In distributed methods, the data is split amongst different administrative units and computations are carried out locally in parallel to each other. The outcome of the local computations are then aggregated into a final result on a central machine.

In this talk we will consider the limitations and guarantees of distributed methods under communication constraints (i.e. only limited amount of bits are allowed to be transmitted between the machines) in context of the many normal means and Gaussian white noise model. We derive minimax lower bounds, matching upper bounds both for testing and estimation.

This is a joint work with Harry van Zanten and Lasse Vuursteen.

Yuhao Wang (Tsinghua University)
Wednesday 9:50am

Title: Long-term Causal Inference Under Persistent Confounding via Data Combination

Abstract: We study the identification and estimation of long-term treatment effects when both experimental and observational data are available. Since the long-term outcome is observed only after a long delay, it is not measured in the experimental data, but only recorded in the observational data. However, both types of data include observations of some short-term outcomes. In this paper, we uniquely tackle the challenge of persistent unmeasured confounders, i.e., some unmeasured confounders that can simultaneously affect the treatment, short-term outcomes and the long-term outcome, noting that they invalidate identification strategies in previous literature. To address this challenge, we exploit the sequential structure of multiple short-term outcomes, and develop three novel identification strategies for the average long-term treatment effect. We further propose three corresponding estimators and prove their asymptotic consistency and asymptotic normality. We finally apply our methods to estimate the effect of a job training program on long-term employment using semi-synthetic data. We numerically show that our proposals outperform existing methods that fail to handle persistent confounders.

Yi Yu (University of Warwick)
Wednesday 3pm

Title: Local differential privacy: User-level and federated learning

Abstract: With decentralised data settings and differential privacy constraints when it comes to communication among servers, we consider two different types of privacy constraints within each server. In the first part of the talk, I will talk about local differential privacy constraints within each server, and this is available at https://arxiv.org/abs/2405.11923. In the second half, I will talk about central differential constraints within each server, and this is available at https://arxiv.org/abs/2403.11343.

Poster abstracts

Alberto Bordino (University of Warwick)

Title: Tests of Missing Completely At Random based on sample covariance matrices

Abstract: We study the problem of testing whether the missing values of a potentially high-dimensional dataset are Missing Completely at Random (MCAR). We relax the problem of testing MCAR to the problem of testing the compatibility of a sequence of covariance matrices, motivated by the fact that this procedure is feasible when the dimension grows with the sample size. Tests of compatibility can be used to test the feasibility of positive semi-definite matrix completion problems with noisy observations, and thus our results may be of independent interest. Our first contributions are to define a natural measure of the incompatibility of a sequence of correlation matrices, which can be characterised as the optimal value of a Semi-definite Programming (SDP) problem, and to establish a key duality result allowing its practical computation and interpretation. By studying the concentration properties of the natural plug-in estimator of this measure, we introduce novel hypothesis tests that we prove have power against all distributions with incompatible covariance matrices. The choice of critical values for our tests rely on a new concentration inequality for the Pearson sample correlation matrix, which may be of interest more widely. By considering key examples of missingness structures, we demonstrate that our procedures are minimax rate optimal in certain cases. We further validate our methodology with numerical simulations that provide evidence of validity and power, even when data are heavy tailed.

Mengchi Chen (University of Warwick)

Title: Doubly robust Bayesian inference for non-linear regression under Berkson measurement error

Abstract: Standard Bayesian inference for regression problems can be sensitive to model misspecification as well as measurement error (ME) of covariates. This project explores the double robustness of Bayesian Nonparametric Learning (NPL) posteriors, aiming to address potential inaccuracies in the regression error model and Berkson ME in the covariates, where ME is independent of the observed variables. Such scenarios are common in practical applications, such as measuring the impact of standard drug doses on disease outcomes or assessing the effects of air pollution on public health via fixed monitoring stations. In these cases, observed data may be subjected to Berkson ME, and the models used might only approximate real-world phenomena, thus often being misspecified. This study seeks to achieve more reliable inferences even when conventional assumptions about data accuracy and model fit are compromised.

Oliver Y. Feng (University of Bath)

Title: Optimal convex M-estimation via score matching

Abstract: In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. Our semiparametric approach targets the best decreasing approximation of the derivative of the log-density of the noise distribution. At the population level, this fitting process is a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. The procedure is computationally efficient, and we prove that our procedure attains the minimal asymptotic covariance among all convex M-estimators. As an example of a non-log-concave setting, for Cauchy errors, the optimal convex loss function is Huber-like, and our procedure yields an asymptotic efficiency greater than 0.87 relative to the oracle maximum likelihood estimator of the regression coefficients that uses knowledge of this error distribution; in this sense, we obtain robustness without sacrificing much efficiency. Numerical experiments using our accompanying R package 'asm' confirm the practical merits of our proposal.

Alexandre Galashov (UCL Gatsby and Google DeepMind)

Title: Deep MMD Gradient Flow without adversarial training

Abstract: We propose a gradient flow procedure for generative modeling by transporting particles from an initial source distribution to a target distribution, where the gradient field on the particles is given by a noise-adaptive Wasserstein Gradient of the Maximum Mean Discrepancy (MMD). The noise-adaptive MMD is trained on data distributions corrupted by increasing levels of noise, obtained via a forward diffusion process, as commonly used in denoising diffusion probabilistic models. The result is a generalization of MMD Gradient Flow, which we call Diffusion-MMD-Gradient Flow or DMMD. The divergence training procedure is related to discriminator training in Generative Adversarial Networks (GAN), but does not require adversarial training. We obtain competitive empirical performance in unconditional image generation on CIFAR10, MNIST, CELEB-A (64 x64) and LSUN Church (64 x 64). Furthermore, we demonstrate the validity of the approach when MMD is replaced by a lower bound on the KL divergence.

Alexander Kent (University of Warwick)

Title: Rate Optimality and Phase Transition for User-Level Local Differential Privacy

Abstract: Most of the literature on differential privacy considers the item-level case where each user has a single observation, but a growing field of interest is that of user-level privacy where each user holds multiple observations and wishes to maintain the privacy of their entire collection. Whilst commonly considered in the central model of differential privacy, there is comparatively less work within the local model.

In this work, we consider the local model of user-level differential privacy where we prove a general minimax lower bound, which shows that, for any locally private user-level estimation problem, the risk cannot be made to vanish for a fixed number of users even when each user holds an arbitrarily large number of observations. We then derive matching lower and upper bounds for univariate and multidimensional mean estimation, sparse mean estimation and non-parametric density estimation. In particular, we observe a phase-transition in the minimax rates when the number of samples each user holds is sufficiently large relative to the number of users.

Further, in the case of (non-sparse) mean estimation and density estimation, we see that, up until the phase transition boundary, the rate is the same as having an equivalent number of users in the item-level setting. However different behaviour is observed in the case of sparse mean estimation wherein consistent estimation is impossible when the dimension exceeds the number of observations in the item-level setting, but is possible in the user-level setting where the number of observations per user exceeds the ambient dimension, which may be of independent interest for applications as an example of a high-dimensional problem that is feasible under local privacy constraints.

Marie Analiz April Limpoco (Hasselt University)

Title: Federated mixed effects logistic regression based on one-time shared summary statistics

Abstract: Upholding data privacy especially in medical research has become tantamount to facing difficulties in accessing individual patient data. Federated learning has emerged as an option to preserve privacy of individual observations while still estimating a global model that can be interpreted on the individual level. It usually involves iterative communication between the data analyst and data providers. Model parameter estimates are updated either through manual transfer or through a computer network which does not require the analyst to have access to individual observations. In this paper, we present a strategy to estimate a binary logistic regression with random intercept which requires data providers to share only summary statistics once. It involves generating pseudo-data that matches the supplied summary statistics and using these into the model estimation process instead of the actual data. Our strategy is able to include multiple predictors which can be a combination of continuous and categorical variables. Through simulation, we show that our approach estimates the true model at least as good as the one which requires the pooled individual observations whenever the sample per data provider is sufficiently large (i.e. $n \geq 250$ for a single data provider ($m = 1$), $n \geq 30$ for $m = 5$, $n \geq 20$ for $m = 10$, $n \geq 10$ for $m = 15, 20$, $n \geq 5$ for $m = 50$, and $n \geq 2$ for $m = 100$). We demonstrate how it works on publicly available. In conclusion, we are able to show that federated logistic regression with random intercept per data provider is possible with only summary statistics shared only once by the data providers to the data analyst. Unlike typical federated learning algorithms, our approach eliminates infrastructure requirements and security issues while being communication efficient and accounting for heterogeneity

Tianyi Ma (University of Cambridge)

Title: Departures from Missing (Completely) at Random: Decision-theoretic foundations and minimax rates

Abstract: We consider the estimation of population parameters—namely the mean and the coefficients of a linear model—in the presence of missing data. Moving beyond restrictive missing completely at random (MCAR) assumptions and motivated by the incompatibility index of Berrett and Samworth (2023), we first formulate a missing data analogue of Huber's ϵ-contamination model. We then show that, for mean estimation, the minimax rate of convergence decomposes into a sum of the minimax rate under a heterogeneous, MCAR assumption and a robust error term incurred by departure from MCAR.

We next re-visit the class of contamination distributions and introduce the finer notion of realisability, which is rich enough to capture all missing not at random distributions constructed from a base distribution P. An infinite-dimensional version of Farkas's lemma is employed to provide a necessary and sufficient condition for a distribution to be realisable. In the context of regression with a missing response, our realisable class captures departures from missing at random (MAR) and generalises well-known conditions in causal inference such as the sensitivity condition of Rosenbaum (1987).

Over realisable classes, we establish minimax rates for both mean estimation and linear regression with a missing response. For univariate mean estimation, and when P is Gaussian, our minimax optimal estimator is consistent; when P admits a finite ψ_r Orlicz norm, for some $r \geq 1$, the sample mean on the observed data is minimax optimal; finally, if P is only known to admit a finite variance, then the median of means estimator is adaptive to ϵ and minimax optimal with rate faster than the Gaussian rate in the unconstrained setting. In each of these cases, our estimators enjoy a high breakdown point of $\epsilon = 1 - o(1)$. In higher dimensions, we generalise our estimators to the special case where each observation is either completely observed or not observed at all and show that they continue to be minimax optimal, with a dimension-independent robust error term.

Manuel M. Müller (University of Cambridge)

Title: Isotonic subgroup selection

Abstract: Given a sample of covariate-response pairs, we consider the subgroup selection problem of identifying a subset of the covariate domain where the regression function exceeds a pre-determined threshold. We introduce a computationally-feasible approach for subgroup selection in the context of multivariate isotonic regression based on martingale tests and multiple testing procedures for logically-structured hypotheses. Our proposed procedure satisfies a non-asymptotic, uniform Type I error rate guarantee with power that attains the minimax optimal rate up to poly-logarithmic factors. Extensions cover classification, isotonic quantile regression and heterogeneous treatment effect settings. Numerical studies on both simulated and real data confirm the practical effectiveness of our proposal, which is implemented in the `R` package `ISS`.

Xuzhi Yang (London School of Economics)

Title: Multiple-output composite quantile regression through an optimal transport lens

Abstract: Composite quantile regression has been used to obtain robust estimators of regression coefficients in linear models with good statistical efficiency. By revealing an intrinsic link between the composite quantile regression loss function and the Wasserstein distance from the residuals to the set of quantiles, we establish a generalization of the composite quantile regression to the multiple-output settings. Theoretical convergence rates of the proposed estimator are derived both under the setting where the additive error possesses only a finite ℓ-th moment (for $\ell > 2$) and where it exhibits a sub-Weibull tail. In doing so, we develop novel techniques for analyzing the M-estimation problem that involves Wasserstein-distance in the loss. Numerical studies confirm the practical effectiveness of our proposed procedure