Modern Challenges in Data Decentralization: Federated Learning, Differential Privacy and Communication Constraints

Workshop One on Data robustness

Abstracts

06–10 July 2026

July 1, 2026

Differentially Private Estimation and Inference for Spatial Autoregressive Models

6 July
10.30 am

Huang Danyang

Renmin University of China, China

Privacy-preserving data analysis has attracted increasing attention in modern statistics. However, how to simultaneously protect both the network structure and the nodal information remains a challenging problem. This paper investigates differentially private estimation and inference for Spatial Autoregressive (SAR) models, aiming to protect the privacy of network structure and individual nodal information. We first derive the minimax lower bounds for differentially private estimators under the SAR model, providing a theoretical benchmark for optimal algorithm design. Building on these insights, we propose a differentially private algorithm for SAR parameter estimation and present a comprehensive analysis of its convergence properties, including examinations of representative cases in specific networks. Additionally, we develop differentially private inference procedures for the SAR model. Empirical validation through simulations and real-world data analysis confirms the reliable finite-sample performance of the proposed methodology.

Collaborative Learning Frameworks for Multi-Site Data Networks

8 July
9 am

Rui Duan

Harvard University, USA

Large-scale collaborations across sites are increasingly enabling the collection of diverse datasets and the development of pre-trained models, creating new opportunities for collaborative learning without requiring data sharing or centralization. However, effectively integrating information across sites remains challenging due to heterogeneity in populations, study designs, and data quality, as well as constraints such as limited sample sizes or missing labels in some settings.

In this talk, I will present methodological frameworks for collaborative learning that leverage information across multiple sites to improve performance in a target population. These approaches are designed to flexibly combine models or data from different sources, accommodating varying levels of supervision and heterogeneity across sites. I will highlight strategies for model transfer, aggregation, and adaptation that enable efficient use of existing resources while preserving privacy.

Applications will illustrate how these methods enhance predictive performance and generalizability in real-world multi-site settings. Overall, this work demonstrates the

potential of collaborative learning to support scalable, robust, and privacy-preserving analysis across diverse populations.

Transfer and Multi-Task Learning: Statistical Insights for Modern Data Challenges

8 July
10.30 am

Yang Feng
New York University, USA

In the era of big data, borrowing strength across related tasks through Transfer Learning (TL) and Multi-task Learning (MTL) is essential for prediction efficiency. However, a major challenge is ensuring adaptive transfer while avoiding "negative transfer" from misleading source data. This talk presents a framework for addressing these challenges across three settings. First, for high-dimensional Generalized Linear Models (GLMs), I introduce TransGLM, a two-step procedure that uses a source-detection algorithm to filter uninformative sources, thereby improving estimation and prediction performance. Second, I extend these ideas to Unsupervised Federated Learning via the FedGrEM algorithm, which addresses the challenges of learning mixture models across heterogeneous clients without sharing raw data. Finally, I move to Representation Learning, where tasks share a low-dimensional linear representation but differ in downstream relationships. I present a framework that adapts to unknown levels of task similarity, ensuring robustness to adversarial attacks and minimax optimality.

Reference: https://yangfengstat.github.io/projects/transfer_learning/

Trustworthy Learning across Heterogeneous Data: Differential Privacy and Adversarial Contamination

6 July
3.30 pm

Mengchu Li
University of Birmingham, UK

Learning from distributed and heterogeneous data is central to modern data science. Recent advances in learning with multi-source data have shown that effectively integrating information across related datasets can significantly improve algorithmic performance. However, heterogeneity across datasets, in terms of sample size, distributional shift, and data quality, poses fundamental challenges in determining the optimal strategy for aggregating information. Moreover, sharing potentially

sensitive information across distributed units raises serious privacy concerns.

In this talk, I will discuss two related projects addressing these challenges. The first studies federated transfer learning under privacy constraints. We introduce a notion of federated differential privacy, which protects each local data set without assuming a trusted central server, and characterise the statistical costs of privacy and heterogeneity across several statistical problems. The second project focuses on robust multi-task learning against adversarial contamination. We show that several existing regularisation-based approaches suffer from a dimension-dependent contamination error and are therefore statistically suboptimal. Motivated by this gap, we develop a computationally efficient filtering-based method that achieves near-optimal statistical performance over a broad range of model parameters.

Efficient Machine Unlearning with Minimax Optimality

7 July
3.30 pm

Sai Li

Tsinghua University, China

There is a growing demand for efficient data removal to comply with regulations like the GDPR and to purge the statistical influence of biased or corrupted data. This has motivated the field of machine unlearning, which aims to eliminate the influence of specific data subsets without the cost of full retraining. In this work, we propose a statistical method for machine unlearning with convex and smooth loss functions. We exemplify its applications in linear models and logistic regressions and establish their convergence guarantees under mild conditions. We also establish the minimax optimality of the proposed algorithm in linear models when only the pre-trained estimate, forget samples, and a small subsample of the remaining data are available. Our results reveal that the estimation error decomposes into an oracle term and an unlearning cost determined by the forget proportion and the forget model bias. We further establish the results in moderate dimensions by considering Ridge-type regularizations. Numerical experiments and real-data applications on Yelp review data and UK Biobank data are conducted to validate the accuracy and computational efficiency of our proposals.

Trans-MA: Sufficiency-principled Transfer Learning via Model Averaging

9 July
2 pm

Huihang Liu

Shanghai University of Finance and Economics, China

Domain aggregation in multi-source transfer learning faces a critical challenge: effectively integrating knowledge from heterogeneous sources while addressing statistical uncertainties. Existing methods rely on restrictive single-similarity assumptions (i.e., individual or combinatorial similarity) and often neglect practical variability, leading to suboptimal performance. To address these limitations, we propose a sufficiency-principled transfer learning framework that systematically balances model averaging and model selection during domain aggregation with unknown informative knowledge. The framework employs a sufficiency principle for quantifying transferable knowledge to eliminate the challenges of spurious correlation and perturbated evaluation. The proposed model averaging algorithms accommodate both individual and combinatorial similarity regimes, and also has privacy-preserving mechanisms. Theoretically, we establish the asymptotic optimality, estimator convergence and asymptotic normality, for multiple source domain linear regression models with diverging parameters. Especially, compared with existing results, we provide enhanced rate of converge for parameter of interest. Empirical validation through extensive simulations and an analysis of Beijing housing rental data demonstrates the statistical superiority of our framework over conventional domain aggregation methods. The proposed methodology extends beyond regression models, offering a generalizable paradigm for transfer learning in statistical decision theory.

Efficient Differentially Private Regression Inference for Longitudinal Data

9 July
9 am

Marco Avella Medina

Columbia University, USA

Differential privacy provides a rigorous framework for releasing statistical analyses while protecting individual-level information. In longitudinal regression, this protection must apply to an individual's entire trajectory, making user-level privacy the natural requirement. We study efficient estimation and inference for fixed-effects longitudinal linear regression under user-level Gaussian differential privacy. We propose private ordinary least squares as well as feasible generalized least squares procedures based on both distributed local regression estimates and pooled estimating equations. We establish finite-sample utility bounds and asymptotic normality under short-range temporal dependence, identifying when privacy noise is asymptotically negligible. Our results characterize regimes where distributed or pooled private regression is

preferable, depending on covariate heterogeneity, panel length, sample size, and coefficient magnitude. We also develop private heteroskedasticity and autocorrelation consistent covariance estimators with finite sample privacy noise corrections for valid confidence intervals. Simulations and a longitudinal data example demonstrate the predicted efficiency gains and near-nominal coverage.

TBA

Short Talk
10 July
10.30 am

Sara Pinciroli
Bocconi University, Italy

Pending

Order-Optimal 1-Bit Mean Estimation

7 July
2 pm

Jonathan Mark Scarlett
National University of Singapore, Singapore

We study the problem of mean estimation under 1-bit communication constraints. We propose a novel adaptive mean estimator based solely on randomized threshold queries, where each 1-bit outcome indicates whether a given sample exceeds a sequentially chosen threshold. Our estimator is (ϵ, δ)-PAC for any distribution with a bounded mean $\mu \in [-\lambda, \lambda]$ and a bounded k-th central moment $\mathbb{E}[|X - \mu|^k] \leq \sigma^k$ for any fixed $k > 1$. Moreover, our sample complexity is order-optimal in all such tail regimes, i.e., for every such k value. For $k \neq 2$, our estimator's sample complexity matches the unquantized minimax lower bounds plus an unavoidable $O(\log(\lambda/\sigma))$ localization cost. For the finite-variance case ($k = 2$), our estimator's sample complexity has an extra multiplicative $O(\log(\sigma/\epsilon))$ penalty, and we establish a novel information-theoretic lower bound showing that this penalty is a fundamental limit of 1-bit quantization. We also establish a significant adaptivity gap: for both threshold queries and more general interval queries, the sample complexity of any non-adaptive estimator must scale linearly with the search space parameter λ/σ, rendering it vastly less sample efficient than our adaptive approach.

TBA

7 July
9 am

Ziteng Sun
Google Research, New York, USA

Pending

Robust Bayesian Inference

10 July
9 am

Botond Szabo
Bocconi University, Italy

We consider generalized Bayesian methods under data contamination and outliers. We show that rescaling the likelihood can provide robustness against outliers, achieving the minimax-optimal contraction rate when the proportion of contaminated data is below a certain threshold. We then derive novel methods with improved, optimal contraction rates even when the contamination proportion is high. Our upper bounds are complemented by matching minimax lower bounds. The theoretical results are illustrated through numerical experiments.

TBA

Short Talk
10 July
11 am

Banerjee Tathagata
National University of Singapore, Singapore

Pending

---

ePTR: A Bridge to DP Estimation

10 July
2 pm

Xin Tong
National University of Singapore, Singapore

Differential privacy (DP) is a rigorous framework that protects the participation of individuals in a dataset by limiting information leakage from released estimators. This creates a challenging setting for statisticians: DP must hold uniformly over all possible datasets, whereas statistical practice often downweights atypical or rare outcomes. The conceptual challenge is especially pronounced in sensitivity analysis—the key quantity

governing the magnitude of DP noise and, consequently, estimator accuracy—because many estimators, including ordinary least squares for linear regression, exhibit markedly higher sensitivity on atypical datasets.

Propose–Test–Release (PTR) is designed to address such cases, but its classical implementation requires computing the exact insensitive set and the dataset's Hellinger distance to that set, both of which are typically intractable. We introduce efficient PTR (ePTR), which replaces the exact insensitive set with a simpler subset and the exact Hellinger distance with a Lipschitz-based lower bound. This flexibility enables substantially simpler DP mechanisms that achieve rate-optimal accuracy in many settings.

Robust Multi-task Learning for Principal Component Analysis

6 July
2 pm

Haolei Weng

Southern University of Science and Technology, China

Principal component analysis (PCA) is a fundamental tool in statistical learning. For data collected from multiple sources, we propose new multi-task learning PCA algorithms that can adaptively leverage the unknown distribution similarities to enhance eigenspace estimation and are robust to data contamination. The first proposed algorithm is shown to achieve (near) minimax optimal rate for the average estimation error. Leveraging a notion of matrix depth, we then develop a second algorithm that is minimax optimal for the maximum estimation error under certain conditions. Numerical studies are provided to demonstrate the effectiveness of the proposed methods.

Observation-Level Watermarking and Detection for Tabular Data

9 July
10.30 am

Bi Xuan

University of Minnesota, USA

With the development of generative Al, watermark techniques have been widely used to detect the authenticity of Al-generated data and protect the rights of users and creators. While it is already well applied in data types including imaging and text data, watermarking tabular data are still under-explored. Existing methods primarily focus on numerical data, leaving discrete, categorical, and mixed data less studied. In this work, we propose STAMP (Single-observation Tabular Attribution and Marking

Procedure), a novel framework for watermarking tabular data that can accommodate and preserve a wide range of distributions. We also develop a corresponding detection mechanism, which can reliably identify watermarks even when the sample size is as small as one. We establish theoretical guarantees for asymptotic consistency and detection accuracy. Finally, through extensive simulation studies and two real-data applications, we demonstrate that the proposed method is effective and robust to subsetting, while maintaining data fidelity and a high detection rate.

Optimal Learning for Fairness-Aware Contextual Bandits

9 July
3.30 pm

Gengyu Xue
University of Warwick, UK

Algorithmic fairness has become a central topic in modern machine learning, and mitigating unfair disparities in treatment between similar individuals is a growing concern in sequential decision-making. In this paper, we systematically study statistical learning in contextual bandits under a global covariate-based Lipschitz fairness constraint, which requires individuals with similar contexts to receive similar policies. We show that Lipschitz fairness reshapes the oracle decision rule by inducing a mixing region near the decision boundary, leading to an explicit and quantifiable cost of fairness. We further establish a canonical representation for the optimal fair policy, reducing fair policy design to a lower-dimensional boundary optimisation problem. This structural characterisation reveals an intrinsic exploration property of Lipschitz fairness, which induces covariate diversity without requiring additional forced exploration and motivates the design of a lazy greedy plug-in algorithm for settings with unknown arm-wise regression coefficients. The algorithm preserves the Lipschitz fairness constraints while achieving sharp regret guarantees. We further prove its minimax optimality through a matching lower bound, showing that Lipschitz fairness introduces an unavoidable dimension-dependent boundary-learning cost. Our theoretical findings are complemented by extensive numerical experiments on both synthetic and real-world datasets, demonstrating the practical effectiveness of the proposed algorithm.

Sample-Efficient and Low-Cost Model-Free Reinforcement Learning

10 July
3.30 pm

Lingzhou Xue

The Pennsylvania State University, USA

Reinforcement learning (RL) provides a general framework for sequential decision-making under uncertainty, and in federated reinforcement learning (FRL), multiple agents collaboratively learn under the coordination of a central server without sharing raw data. Recently, we have developed new methodological and theoretical results for model-free RL in tabular episodic Markov Decision Processes across both single-agent and federated settings. In particular, we have established the first gap-dependent regret for federated RL and developed a novel fine-grained analytical framework that yields the fine-grained regret bound. Further, we have proposed new algorithms with provable guarantees for low-cost RL.

Double Robustness vs. Double Flexibility in Unsupervised Domain Adaptation

7 July
10.30 am

Jiwei Zhao

University of Wisconsin, USA

Unsupervised domain adaptation seeks to transfer knowledge from a labeled source domain to an unlabeled target domain by addressing distributional discrepancies. In the literature, two key assumptions describe how these distributions differ: covariate shift and label shift. In this talk, I will introduce two related but distinct concepts under these assumptions: double robustness under covariate shift and double flexibility under label shift. Using techniques from semiparametric statistics, I will highlight their intrinsic similarities. Our findings shed light on the strengths and limitations of each paradigm, offering guidance for future research in robust and flexible domain adaptation strategies.
