Jinjin Tian

Email: jinjint@andrew.cmu.edu | Phone: (412)-616-6460 | Homepage

Education Background

PhD in Statistics and Data Science, Carnegie Mellon University *GPA 4.05 / 4.30 expected May 2023*

- Thesis topic: Higher-order higher-resolution methods in statistical genetics
- Advisors: Jing Lei and Kathryn Roeder

Bachelor in Statistics, University of Science and Technology of China *GPA 3.86 / 4.30, July 2018*

Qualification

- **5+ Years research experience in statistics and machine learning**
- **5+ Years of programming experience, familiar with Python (PyTorch, Tensorflow), R, MATLAB, SQL**

Work Experience

- **Applied Scientist Intern, AWS, Amazon.** (Paper, code) *June – Aug. 2020*
  Under supervision of Lenon Minorics, Professor Guido Imbens.
  - Constructed a new online experimentation algorithm that allows an efficient trade-off between revenue constrain and time constraint in E-commerce, utilizing Multi-Armed Bandits and any-time valid inference.
  - Proposed new algorithm to deal with unknown abrupt changes over time using Thompson Sampling and sequential change detector, which only cost O(1) for each update, and has much lower regret comparing with other state-of-arts.

Research Experience

My research lie in high-dimensional data analysis with special focus on **feature interaction** and **local structure**, as well as error control in **online experimentation**. See Google Scholar for part of publications.

- **Signals recovery in noisy high-dim mixture via local structure learning** *Oct. 2020 – Present*
  Under supervision of Professor Jing Lei & Kathryn Roeder.
  - Proposed a statistic to capture local structure between a pair of features. Proved exact recovery of sparse signals based on this statistics in a high-dim mixture model, under even impossible scenarios for canonical methods like sparse PCA. Working on application in subpopulation detection and dynamic community recovery in genetics.
- **Self-supervised learning with and without contrastive pairs** *Feb. 2021 – Present*
  Under supervision of Professor Andrej Risteski.
  - Investigated the role of augmentation, adversarial samples and network structure in self-supervised learning (SSL) of image, through downstream classification task, feature visualization and texture-shape bias analysis. Working on a sparse coding model to theoretical further explore the differences in feature learning process.
- **Detecting non-linear non-monotone relationship by sample specific dependency** *Feb. 2021 – June. 2021*
  Under supervision of Professor Jing Lei & Kathryn Roeder.
  - Developed a new univariate dependency measure, together with a consistent, tuning-free, nonparametric estimator. Proved its ability to detect non-linear, non-monotone relationship and robustness to outliners. When applied as an independence test, it is competitive among state of the art like taustar, distance correlation, and HSIC.
- **Single-cell gene expression data modeling incorporating gene interaction** (Bioinformatics 2021, package)
  Under supervision of Professor Kathryn Roeder & Jiebiao Wang. *July. 2019 – May. 2020*
  - Modeled the single-cell gene expression data through a hierarchical model to characterize various types of heterogeneity, together with a copula approach to depict higher-order gene interaction. Wrapped up the model as a simulation software which also allows mimicking real data via fitting the model using a moment matching method.
- **Large-scale post-hoc model selection and inference under dependence** (Paper) *Sept. 2019 – May. 2020*
  Under supervision of Professor Aaditya Ramdas & Eugene Katsevich & Jelle Geoman.
  - Developed a class of simultaneous inference methods under arbitrary dependence with only linear time computation. Derived tighter calibration and the corresponding asymptotic power under a Gaussian dependence model.
- **Adaptive algorithms for online error control** (NeurIPS 2019, code) (SMMR 2021, code) *Jan – Sept. 2019*
  Under supervision of Professor Aaditya Ramdas.
  - Constructed new algorithm for online error control via adapting to both signal proportion and noise level, which outperforms current state of arts in terms of applicable range as well as power. Formally proved power gains for some specific error choice, and derived closed-form optima of hyperparameters in a Gaussian sequence model.

Coursework

- **Machine Learning:** Introduction to Advanced ML (A+), Statistical ML (A+), Convex Optimization, Deep Reinforcement Learning, Advanced Deep Learning (A+)
- **Statistics:** Measure Theory, Intermediate Statistics (A+), Advanced Probability Theory, Reproducibility, Regression Analysis (A+), Advanced Data Analysis, Stochastic Process, Time Series
- **Math:** Functional Analysis (A+), Real Analysis (A+), Complex Analysis, Mathematical Analysis, Linear Algebra (A+)
- **Programming:** Statistical Computing, Data structure and Data base, C and Algorithm.