Siddhaarth Sarkar

sid.sarkar8(at)gmail(dot)com | sidsarkar8 | sidsarkar8.github.io

I am a Ph.D. student at Carnegie Mellon University, with research interests in statistics, machine learning, and probability. My primary focus is developing statistical methods for uncertainty quantification that are applicable in diverse contexts.

Education

Statistics and Data Science — *Ph.D.* Aug 2020 - May 2025
Carnegie Mellon University, Pittsburgh, PA — *Advised by Dr. Arun Kuchibhotla*

Statistics — *Master of Statistics* Jul 2018 - May 2020
Indian Statistical Institute, Kolkata — *Theoretical Statistics Specialization*

Statistics — *Bachelor of Science* Jul 2015 - May 2018
Indian Statistical Institute, Kolkata

Skills

Programming Skills
R, Python, SQL, TensorFlow, MATLAB, LaTeX, Git

Technical Skills
Statistical Modeling, Data Analysis, Data Visualization, Machine Learning, Supervised and Unsupervised Learning Algorithms, Optimization Algorithms, Deep Learning

Foundational Skills
Collaborated with researchers across different hierarchies: professors, graduate students, and undergraduate students. Taught and advised students as a teaching assistant on various subjects. Active organizer and participant in departmental and cultural events.

Projects

Application of the post-selection conformal method: police officer safety data Jan 2024 - Present
Keywords: *Police officer risks — Conformal Inference —Post-selection Inference— Black-Box Methods*

- Police officers in the USA face significant risks during service calls, making accurate risk assessments crucial for their safety.
- Black-box machine learning models improve risk assessment by identifying patterns in call data.
- Conformal inference procedures provide prediction sets with target coverage probability guarantees but traditionally require fixed coverage levels.
- The aim is to apply a conformal inference procedure with a post-selection guarantee for coverage levels allowing adjustable coverage levels while maintaining the usual guarantees.
- This approach enhances risk forecasts, improving officer safety during service calls.

Unified Framework for Inference Using Confidence Sets for the CDF Aug 2023 - Present
Keywords: *Confidence Intervals — Assumption-lean Inference — Semi-infinite Optimization — Non-parametric Methods*

- Traditional statistical inference methods often face limitations due to strict assumptions. Methods are typically tailored to specific assumptions, restricting their adaptability.
- Developed a unified framework for deriving confidence intervals for various functionals (e.g., mean or median) under a broad class of user-specified assumptions (e.g., finite variance or tail behavior).
- Leveraged confidence sets for cumulative distribution functions (CDFs) to offer:
  - A principled and flexible inference strategy
  - Reduced dependence on stringent assumptions
  - Applicability in diverse contexts

New Asymptotic Limit Theory and Inference for Monotone Regression Apr 2023 - Dec 2023
Keywords: *Monotonic Regression — Confidence Intervals — Asymptotics*

- Explored and documented interesting properties of monotonic least squares estimates, including rate of convergence, adaptivity properties, and pointwise asymptotic distribution.
- Investigated the full richness of asymptotic limit theory in nonparametric regression.
- Developed and validated an asymptotically valid confidence interval method given the new class of asymptotic limits.

Post-selection Inference for Conformal Prediction: Trading Coverage for Precision Oct 2022 - Present
Keywords: *Conformal Inference — Black-Box Methods — Reproducibility — Distribution-free Inference*

- Traditionally, conformal prediction inference requires a data-independent specification of miscoverage level. Practical applications often require updating the miscoverage level after computing the prediction set.
- Developed a *simultaneous conformal inference* procedure to account for data-dependent miscoverage levels. This allows practitioners to trade coverage probability for prediction set quality while maintaining statistical validity.

Consistency and Inference for Density Estimation Trees Jul 2022 - Present
Keywords: *Random Forests — Consistency Theory — Greedy Algorithms — Density Estimation*

- Density estimation trees (DETs) are a data-driven greedy partitioning procedure that returns a tree-structured density estimator. Theoretical behavior of DETs is challenging to understand due to its greedy approach.
- Proved consistency results for DETs under general regularity conditions on the target density. These conditions apply universally across a broad class of distances (Bregman divergences).
- Improved theoretical understanding and practical applications of DETs for density estimation.

Spatio-temporal Methods for Inferring Trajectories of Under-ice Argo Floats Dec 2020 - Jan 2022
Keywords: *State-space Modeling — Kalman Filters — Data Imputation — Uncertainty Quantification*

- The Argo program uses an array of autonomous profiling floats equipped with sensors and tracked via GPS, drifting along ocean currents to collect data on key oceanographic variables.
- When these floats are under sea ice, their inability to reach the surface results in a loss of location data, critical for scientific analysis with Argo data.
- Proposed a state-space modeling approach to infer the missing trajectory of an under-ice float using an efficient Kalman smoother algorithm that leverages temperature and salinity information.
- Improved the quantification of the uncertainty of predicted floating locations, enhancing the reliability of downstream scientific data usage.

Design of Experiments for Peer Effects May 2019 - Dec 2020
Keywords: *Causal Inference — Networks — Interference — Mixed-integer Optimization*

- Modern randomized experiments in political science, sociology, online marketing, and health sciences study not only direct causal effects but also how the treatment of one unit could affect the outcome of another.
- Network information can provide an important description of potential interference patterns in a causal problem.
- Designed a randomized design scheme derived from a binary linear program, allowing estimation of peer influence parameters. Explored simultaneous identifiability issues and arbitrary neighborhood interference function estimation.

Publications and preprints

[1] Mallick S, Sarkar S, Kuchibhotla AK. New Asymptotic Limit Theory and Inference for Monotone Regression. arXiv preprint arXiv:231020058. 2023.

[2] Sarkar S, Kuchibhotla AK. Post-selection Inference for Conformal Prediction: Trading off Coverage for Precision. arXiv preprint arXiv:230406158. 2023.

Awards

D. Basu Memorial Gold Medal Award Aug 2019

- For outstanding seminar as well as best performance in B.Stat. Program.
- Title of seminar: ***False discovery rates: A powerful multiple testing tool***

KVPY Scholarship May 2015

- Funded by **Department of Science and Technology, Govt. of India**, aimed at encouraging students to take up research careers in the areas of basic sciences.

Invited Talks

Unified Framework for Inference Using Confidence Sets for the CDF Aug 2024
Joint Statistical Meetings, Portland

- Contributed papers in "New Methods for Valid and Flexible Inference".

Post-selection Inference for Conformal Prediction Dec 2023
International Conference of the ERCIM WG on Computational and Methodological Statistics, Berlin

- Contributed papers in "Nonparametric Inference and Decision Making".

Post-selection Inference for Conformal Prediction Aug 2023
Joint Statistical Meetings, Toronto

- Contributed papers in "Concentration and Conformal Prediction".

Design of Experiments for the Identification and Estimation of Peer Effects Dec 2019
International Indian Statistical Association Conference, Mumbai
- Student poster presentation.

Teaching Experience

Teaching Assistant — *Carnegie Mellon University* Aug 2020 - Present
- Introduction to Statistical Inference
- Intermediate Statistics
- The ABCDE of Statistical Methods in Machine Learning
- Advanced Methods for Data Analysis
- Probability Theory for Computer Scientists
- Engineering Statistics and Quality Control
- Regression Analysis
- Multilevel and Hierarchical Models

Service

Reviewer
NeurIPS (2023), Annals of Statistics (2023-24)

Pittsburgh Bengali Student Association — *Cultural Committee Member* Aug 2022 - Present
- Organized various cultural events for Bengali students in Pittsburgh as a key member of the Bengali Student Association.

References

Arun Kuchibhotla Associate Professor of Statistics and Data Science
Carnegie Mellon University
Email: arunku@stat.cmu.edu

Richard Berk Professor of Criminology and Statistics
University of Pennsylvania
Email: berkr@sas.upenn.edu

Alessandro Rinaldo Associate Professor of Statistics
University of Texas Austin
Email: alessandro.rinaldo@austin.utexas.edu