Last Updated on November 6, 2025

LONGTIAN SHI

+86 15639073546 ◇ Shenzhen, Guangdong, China
shilt2022@mail.sustech.edu.cn; LinkedIn; Personal Website

EDUCATION

Southern University of Science and Technology (SUSTech), China **SEP 2022 - Expected JUN 2026**

- **Major**: **Statistics** (Advised by Chair ProfessorQi-Man Shao). **Major GPA: 3.96/4.00**.
- Overall GPA: **3.91/4.00 (2/44)**. Courses: Topics in Probability and Statistics (100, PhD-level measure-based), Statistical Learning (100), Python (97), Time Series Analysis (99), Mathematical Statistics (99), Statistical Linear Model (98), Nonparametric Statistics (97), Bayesian Statistics (95), Advanced Linear Algebra (90)
- Minor: Finance (Minor GPA: 3.88/4.00). Financial Investment (96), Marketing (95), Economics (94), Management Information System (94), Advanced Operations Research.

Summer Session, University of California, Davis **JUN 2025 - AUG 2025**
Courses: ECN 140 Econometrics and MAT 127C Real Analysis (100, A+)

PUBLICATIONS&MANUSCRIPTS

1. **Shi, L.**, Shi, Y., Fu, Y., Jiang, F., Ma, Y. (2025+). ***The Prize Premium in Publishing Timelines***. Under Review of the *Journal of Informetrics (JOI)*.
2. Zhang, X.†, **Shi, L.**†, Zhao, H. (2025+). ***A Novel Empirical Bayes Method for Genetic Fine-mapping with GWAS Summary Statistics***. Manuscript to be submitted.
3. **Shi, L.**†, Chu, H.†, Zhou, D., Liu, M. (2025+). ***Kernel-assisted Debiased Inference of Surrogate Model Predictive Metrics under Pairwise Covariate Shifts***. Manuscript in preparation.

†These authors contributed equally to this work.

RESEARCH EXPERIENCE

Independent Research on Causal Machine Learning and Robust Inference **FEB 2025 - PRES**

- Supervised by Assistant Professor Molei Liu, Peking University. Advised by Prof. Tianxi Cai, Harvard.
- I am responsible for **theoretical derivation, simulation investigation, and real data application into how kernel smoothing could be used in statistical inference of debiased/double machine learning**. Parameters of interest include TPR, FPR, ROC, and AUC, which accommodate three data sources: human-labeled (gold standard), AI-labeled surrogates, and an unlabeled target set, under various conditions, including pairwise covariate shifts. By approximating an indicator function with a regularized kernel function and employing debiasing techniques, such as Neyman Orthogonality and cross-fitting, we successfully demonstrated that the debiased cross-fitting estimators are $\sqrt{n}$-consistent under the transfer learning setting with multiple covariate shifts between datasets. The doubly robustness is also illustrated theoretically. In the simulation, the density ratio is estimated based on posterior probabilities after splitting the sample. Already validated in real-world datasets like MIMIC-III and -IV, our framework rigorously addresses AI-driven problems, such as evaluating the predictive performance of AI tools as surrogates for gold-standard labels in Electronic Health Records (EHR).

Research on Causality in Computational Social Science and Networks **AUG 2023 - APR 2025**

- I led a research project under the supervision of Assistant Professor Yifang Ma at the Department of Statistics and Data Science in SUSTech. I also regularly attended Prof. Ma's seminar on network science and computational social science, starting in September 2023.

- Leveraging a large-scale, multi-source dataset linking OpenAlex and PubMed, we assembled 1,168,808 publication records organized into 1,778 winner–coauthor groups and designed quasi-experimental methods (fixed-effects regressions and difference-in-differences event studies) to identify **causal effects** of the winner's prestige on submission-to-acceptance time. Our findings show that prize winners experience a 7-12 day reduction in acceptance time in top journals (17-30 days in elite venues, such as *Nature*), with advantages peaking around the award year and then decaying but remaining significant. We also discovered more substantial impacts for younger academics and in high-discretion environments. This manuscript represents my first **first-author** work and is currently under review at the *JOI*, and another manuscript is presently in preparation.

Research on Empirical Bayes Methods for Genetic Fine-mapping **JUN 2024 - SEP 2024**

- Supervised by Prof. Hongyu Zhao, Department of Biostatistics, Yale University. On-site internship.
- I am responsible for developing **E**mpirical-**B**ayes for **F**ine **M**apping, namely **EBFM**, a biostatistical method for enhancing identification of disease-causing genetic features using GWAS summary statistics. The proposed method EBFM utilizes the **spike-and-slab prior and posterior maximization** for estimating the genetic architecture and then captures the Credible Sets of SNPs based on the posterior inclusion probability via **greedy search**. The greedy search is constructed based on the correlation score and is adjusted accordingly in this particular setting. EBFM is more powerful with a lower FDR (False Discovery Rate) by capturing more credible sets with fewer SNPs in each of them, yielding a higher replication rate, precision-recall rate, reproduction rate, etc., in both simulation and real-data studies via the data of European's and African's BMI, UK Biobank, and 1KGP. The simulation and real-data applications (BMI data from different ancestries) aim to compare EBFM's performance with that of existing popular methods, such as SuSiE and CARMA.

Project on Applying Statistical Learning Methods on sedaDNA Datasets **DEC 2024 - PRES**

- Led by Prof. Rasmus Nielsen at the Department of Statistics, University of California, Berkeley.
- I am responsible for implementing high-dimensional sparse PCA, multiple association testing correction, and novel regularized (sparse) regression methods, including uniLasso, on sedaDNA (Sedimentary Ancient DNA) allele frequency data and the environmental metadata, working to uncover evolutionary patterns and to predict future environmental changes. The population bias is corrected by including the principal components in the regularization methods. After matching the selected SNPs to the organisms from a mapping table constructed via read and accession IDs in the BAM files and the NCBI datasets, several species, primarily bacteria, were found whose genetic variants could likely be attributed to environmental features like the mean annual temperature. We plan to apply the pipeline on a much larger dataset and publish our results in *Science*.

SKILLS&LANGUAGES

Coding: Python&R (Specialized), SQL (Proficient); Competent in LaTeX, STATA and Linux
English: Fluent, TOEFL (109, Speaking 24), GRE (Verbal 155 + Quant 170 + Writing 4.0); Chinese: Native

AWARDS&COMPETITIONS&HONORS&MISCELLANEOUS

China National Scholarship ($< 0.4\%$) 2025
Guo Xie Birong Scholarship Excellence Award ($< 1\%$) 2025
Candidate for Student of the Year (6 out of over 5,000 undergraduates) 2025
First Prize (<1%) of University Merit Student Scholarship 2023&2024&2025
Gold Medal for iGEM (International Genetically Engineered Machine) Competition 2024
Second Prize of The Chinese Mathematics Competitions 2023
Second Prize of China Undergraduate Mathematical Contest in Modelling 2023
University Top 10 Volunteer Candidate (Annually Over 100 Hours of Volunteering) 2023
President of the Students' Union and College Peer Tutor of Shuli College MAR 2023 - SEP 2024
The First Level Athletic of Land Rowing in China SEP 2024
Member of College Basketball Team