LONGTIAN SHI

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

EDUCATION

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

- **Major**: **Statistics** (Advised by Chair Professor Qi-Man Shao). **Major GPA: 3.96/4.00**.
- Overall GPA: 3.91/4.00 (2/44). Selected courses: Topics in Probability and Statistics (100, measure-based graduate-level), Statistical Learning (100), Introduction to Python Programming (97), Time Series Analysis (99), Mathematical Statistics (99), Statistical Linear Model (98), Nonparametric Statistics (97), Bayesian Statistics (95), Advanced Linear Algebra (90), Discrete Mathematics and Its Applications (96).
- Minor: Finance (Minor GPA: 3.90/4.00). AI and Its Applications in Finance (100), Financial Investment (96), Corporate Finance (96), Economics (94), Management Information System (94).

Summer Session, University of California, Davis **JUN 2025 - AUG 2025**
GPA: 4.0/4.0. Mathematics and Statistics. Courses: Econometrics and Real Analysis (100, A+).

PUBLICATIONS&MANUSCRIPTS

1. **Shi, L.**, Shi, Y., Fu, Y., Jiang, F., & Ma, Y. (2025). ***The Prize Premium in Publishing Timelines***. Forthcoming in the *Journal of Informetrics*.
2. Zhang, X.†, **Shi, L.**†, & Zhao, H. (2025). ***A Novel Empirical Bayes Method for Genetic Fine-mapping with GWAS Summary Statistics***. (Manuscript).
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).
   †Co-first authors.

RESEARCH EXPERIENCE

Independent Research on Causal Machine Learning and Robust Inference **FEB 2025 - PRES**
Supervised by Assistant Professor Molei Liu, Peking University. Advised by Professor Tianxi Cai, Harvard.

- I am responsible for the theoretical derivation, simulation investigations, and real data applications into how kernel smoothing could be used in the robust inference for **debiased/double machine learning**. Parameters of interest include TPR, ROC, and AUC, which accommodate three data sources: human-labeled (gold standard), AI-labeled surrogates, and an unlabeled target set, under pairwise covariate shifts.
- By approximating the indicator function with a regularized kernel function and employing debiasing techniques (e.g., Neyman Orthogonality and cross-fitting), we demonstrated that the debiased cross-fitting estimators are $\sqrt{nh}$-consistent under this transfer learning setting. The double robustness and $\sqrt{n}$-consistency in the estimation of prevalence were also examined theoretically. In the simulation, the density ratio was estimated based on posterior probabilities after splitting and relabelling the samples.
- Already validated in Electronic Health Records (EHR), MIMIC-III and MIMIC-IV, our framework corrects the AI-induced bias and kernel approximation error, evaluating the predictive performance of AI models as surrogates for human labels in EHR. We aim for top-tier journals like *JASA*, and I will be the co-first author.

Research on Causality in Computational Social Science and Networks **AUG 2023 - APR 2025**
I led a research project under the supervision of Associate Professor Yifang Ma at the Department of Statistics and Data Science in SUSTech, and I regularly attended Prof. Ma's research seminar on network science.

- Leveraging a large-scale (1,168,808 publication records) dataset linking OpenAlex and PubMed, we assembled winner–coauthor groups and designed quasi-experimental methods, including fixed-effects regressions and DID event studies, to identify **causal effects** of the prize winner's prestige on submission-to-acceptance time.
- Our findings include that winners experience a 7-12 day reduction in acceptance time in Nature Index 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 represents my first **first-author** work and will be published at the *Journal of Infometrics*, and another manuscript is presently in preparation.

Research on Empirical Bayes Methods for Genetic Fine-mapping **JUN 2024 - SEP 2024**
Supervised by Professor Hongyu Zhao, Department of Biostatistics, Yale University. On-site internship.

- We developed **E**mpirical **B**ayes for **F**ine **M**apping (**EBFM**), a novel 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 20% more powerful with a lower 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 across 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 Professor Rasmus Nielsen at the Department of Statistics, University of California, Berkeley.

- I implemented 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.
- After matching the selected SNPs to the organisms through a mapping table (constructed via read and accession IDs in the BAM files and the NCBI datasets), several species were found whose genetic variants were associated with environmental features like the mean annual temperature. We are implementing the pipeline on much larger datasets and will publish our results in top-tier journals like *Science*.

SKILLS

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&TEACHING&MISCELLANEOUS

(China) National Scholarship ($< 0.4\%$)	2025
Candidate for Student of the Year (6 out of over 5,000 undergraduates)	2025
Guo Xie Birong Scholarship Excellence Award ($< 1\%$)	2025
First Prize ($< 1\%$) of University Merit Student Scholarship	2023&2024&2025
Tutorial Teaching Assistant for 'Foundation of Probability Theory'	Fall 2025
Gold Medal of International Genetically Engineered Machine (iGEM) Competition	2024
Second Prize of The Chinese Mathematics Competition	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	
Member of College Basketball Team	