Shuting Shen

Mailing Address: 11 Peabody Terrace, Cambridge, MA 02138, Unit 608

+1 8577563165 shs145@g.harvard.edu

Education Background

Department of Biostatistics, Harvard University, Boston, United States 08/2018 - Present

- PhD in Biostatistics, GPA: 3.931
- Courses: Inference I, Inference II, Method I, Data Structures and Algorithms, Probability II, Multivariate Statistical Analysis, Advanced Regression, Optimization, Discrete Probability, Bayesian Methodology in Biostatistics
- Awards: Robert B. Reed Prize (awarded each year to the student(s) receiving the highest grade on the Department's written qualifying exam.)

School of Foundational Education, Peking University Health Science Center, Beijing, China 09/2013-07/2018

- Bachelor in Biomedical English, Major GPA: 90.4/100, Ranking: 1/36
- Core courses: Biostatistics (96/100), Epidemiology (89/100), General Biology (96/100), Organic Chemistry (97/100), Advanced Mathematics (92/100), Biochemistry (95/100), Histology and Embryology (99/100), Physiology (97/100), Immunology (98/100).
- Awards: Special Grade Scholarship (top 2% in 2016); First Grade Scholarship (top 4% in 2015 and 2017); Merit Student Awards at PKU (top 10% in 2015, 2016 and 2017)

School of Mathematical Sciences, Peking University (PKU), Beijing, China 09/2015-07/2018

- Bachelor in Mathematics, Major GPA: 95.9/100, Ranking: top 1%
- Core courses: Mathematical Analysis (99/100), Advanced Algebra (90/100), Functions of Real Variables and Functional Analysis (99/100), Abstract Algebra (94/100), Probability (93/100), Theory of Functions of a Complex Variable (98/100), Statistics (97/100), Ordinary Differential Equations (97/100).

Publication

- Preprint: Shen, S., and Lu, J. (2020).'Combinatorial-Probabilistic Trade-Off: Community Properties Test in the Stochastic Block Models', *arXiv preprint arXiv:2010.15063*

Research Experience

Distributed Fast Principle Component Analysis for Large-scale Data 06/2020 – Present

Advisor: ***Xihong Lin****, Professor, Department of Biostatistics, Harvard University*

- Develop a scalable distributed PCA method, that leverages FAST PCA and distributed estimation of principal eigenspaces, for big data with both large sample size and high dimensionality.
- Derived theoretical error bound for the proposed method that shows the method enjoys the same error rate as traditional full sample PCA.
- Characterized the asymptotic normality of the estimator and its inferential applications.
- Conducted extensive simulations comparing the algorithm with existing methods, showing the high computational efficiency and statistical accuracy of the proposed method.
- Implemented the algorithm on the 1000 Genomes Data.
- Finished the manuscript to be submitted to *JASA-T&M.*

Longitudinal Emotional Analysis on the How We Feel Data 06/2020 – Present

*Advisor: **Xihong Lin**, Professor, Department of Biostatistics, Harvard University*

- Explored the factors that have been influencing the emotional status of US residents since the pandemic.
- Cleaned the dataset using R.
- Conducted GEE logistic regression to analyze the missing pattern.
- Conducted GEE regression with and without IPW correction to analyze factors influencing the emotional status of the US users of the How We Feel app, which surveys COVID symptoms and health status.
- Finished the manuscript to be submitted to *Nature Human Behavior*.

Inference on Choice Model through Markov Chain Approximation 10/2021 - Present

*Advisor: **Junwei Lu**, Assistant Professor, Department of Biostatistics, Harvard University*

- Simplified the assortment optimization algorithm under the Generalized attraction model (GAM)
- Set up outline for the inferential framework
- Working on the theoretical details for distributional characterization of the test statistic

Hypothesis Testing for Clustering Properties Based on Likelihood Ratio Test 01/2019 – 10/2020

*Advisor: **Junwei Lu**, Assistant Professor, Department of Biostatistics, Harvard University*

- Developed a general inference method for clustering property tests based on the log likelihood ratio statistic in Homogeneous Stochastic Block Model.
- Proved the validity of the test in controlling type 1 and type 2 error rates.
- Provided a general lower bound for the clustering property test with side results on exact recovery rate.
- Implemented numerical experiments on both the synthetic data and the protein interaction application to show the validity of our method.
- Finished the manuscript "Combinatorial-Probabilistic Trade-Off: Community Properties Test in the Stochastic Block Models", submitted to *JASA-T&M*.

Estimating the Proportion of Disease Heritability Mediated by Gene Expression Levels 07/2017-10/2017

*Research Assistant | Advisor: **Hongyu Zhao**, Chair, Department of Biostatistics, Yale School of Public Health*

- Presentations for journal club on published paper to share latest progress in the field of genome-wide association study.
- Utilized Matlab to design and perform simulations to assess the robustness of a method proposed in a published essay for estimating the proportion of trait heritability mediated by gene expression levels in given tissues.
- Conducted analysis on the results of simulations to elucidate the sources of bias in estimation results.
- Applied the author's methods to real data from dataset GTEx and GWAS summary statistics and obtained estimation results for 77 traits in 47 tissues.

Biological Big Data: Analysis, Calculation, and Prediction 01/2016-02/2016

*Research Assistant | Advisor: **CHEN Xing**, Assistant Professor at Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

- Completed four training projects of programming that respectively implemented the algorithms proposed by four articles on the prediction of human disease-related risk factors.
- Independently developed an improved random walk on the heterogeneous network for the prediction of

potential miRNA-disease associations.

- Successfully utilized Matlab to implement the random walk and produced prediction results.

Skills and Others

Computer skills: R language, Python, Latex, Matlab, C language, SAS, Stata