

Weijia Xiong

100 Haven Avenue, Tower 3, Apt. 9H, New York, NY 10032
(332)207-5898/(+86)15900930235 | wx2233@cumc.columbia.edu

EDUCATION

Columbia University, Mailman School of Public Health

New York, NY

Master of Science, Biostatistics GPA: 3.92

expected May 2021

- **Relevant course:** Probability, Biostatistics Method, Statistical Inference, Causal Inference, Survival Analysis, Clinical Trial Methods, Epidemiology, Data Science, Statistical Computing, Data mining

East China Normal University

Shanghai, China

Bachelor of Science, Statistics GPA: 3.6

June, 2019

University of California at Los Angeles

Los Angeles, CA

Summer Program, Statistics

Summer, 2017

PROFESSIONAL EXPERIENCE

Roche

Shanghai, China

Statistical Programming Analysis (SPA) Intern

July, 2018 – Dec, 2018

- Participated in development of AutoReport RShiny app to streamline regulatory reporting process and integrate all related data files and documents in a one-stop shop; this app, evaluated highly by department global leader, is currently online and actively used
- Completed design of GitHub web pages for team projects
- Assisted team for with user acceptance testing (UAT), including questionnaire creation, data collection and review

Shanghai Shucui Big Data Technology Co., Ltd.

Shanghai, China

Course Construction Team Member

Apr, 2018 – Nov, 2018

- Cooperated with team members to finish preliminary statistical analysis, time series analysis, course design, and courseware preparation for Principle and Application of Big Data Technology
- Used R to generate html and pdf courseware involving theoretical statistical knowledge and real case analysis

RESEARCH EXPERIENCE

Columbia University Research

New York, NY

• Exploration of Super Learner algorithm

Aug, 2020 – Present

Advisor: Dr. Caleb Miles

- Learned and reproduced the analysis of OTNY data, predicting three months education enrollment or employment among unenrolled and unemployed at baseline via Super Learner algorithm
- Applied Super Learner to Heart disease data set and compared the predicted classification error rate with other machine learning method(eg. GLM, Lasso, SVM, etc)

• An application of Hidden Markov model for mobile mental health study

June, 2020 – Present

Advisor: Dr. Linda Valeri

- Conducted research on evaluation of the performance of Hidden Markov Model for mental health regarding mobile-communication data
- Simulated outcome survey data and covariates (eg. mobile-communication data like phones and calls, season, weekends, monthly clinical diagnostic score) with mutual correlations (eg. AR(1), nonlinear) between latent disease states
- Performed the Hidden Markov model and explored the causal effect of mobile-communication data
- Re-performed the Hidden Markov model for different missing data scenarios(MCAR,MAR,MNAR) with imputation

• Longitudinal microbiome data analysis

Apr, 2020 – Oct, 2020

Advisor: Dr. Jianhua Hu and Dr. John Spivack

- Participated in longitudinal microbiome research. Conducted literature search and assisted in real vaginal microbiome data analysis

- Imputed missing values via KNN method
- Reproduced the clustering work for vaginal community state type and performed different clustering methods (Hierarchical clustering, K-means)
- Conducted comparison of preterm birth classification for different methods(Lasso, Random Forest, SVM, etc)
- **COVID-19 analysis via logistic curve and clustering** Apr, 2020 – May, 2020
 - Analyzed the growth rate, maximum number of cases and mid-point for regions through a logistic curve with gradient decent optimization and calculating start values through ordinary differential equation (ODE)
 - Clustered the countries by K-mean and Gaussian mixture model with EM algorithm
 - Visualization through world maps and 3D plots to compare these two methods
- **Analysis of vehicle collision in New York City** Sep, 2019 – Dec, 2019
Advisor: Dr. Jeff Goldsmith
 - Explored and analyzed NYPD motor vehicle collisions-crashes data and other related covariates dataset such as weather, traffic light, restaurant and holiday
 - Built the regression model and analyzed the correlation between potential factors and vehicle collisions
 - Created a website with vehicle collision map and shiny dashboard for data visualization and conclusions

Department of Statistics, Shanghai Jiao Tong University

Shanghai, China

- **One-round communication efficient distributed M-estimation** Aug, 2020 – Present
 - Proposed one-round communication efficient distributed M-estimation algorithm to obtain CASE(communication-efficient aggregate score equation)estimator and PEN-CASE estimator as for solving lasso problem
 - Implemented algorithm for experiments using Matlab, including both synthetic and real dataset(MNIST and w8a datasets)
- **Byzantine-Tolerant distributed multiclass sparse discriminant Analysis** Apr, 2020 – Aug, 2020
Advisor: Dr. Weidong Liu and Dr. Xiaojun Mao
 - Conducted literature search of LDA and distributed machine learning and wrote and literature review
 - Implemented the Mean-MSDA and Median-MSDA algorithm for experiments using Matlab and R, including both synthetic and real dataset(Heart Disease dataset)
 - Tabulated the results of misclassification rate and draw the plot, wrote analysis of real data

Department of Statistics, East China Normal University

Shanghai, China

- **An application of MS-EGARCH for fluctuation of currency exchange rate** May, 2017 – May, 2018
Advisor: Dr. Jiaqin Wei
 - Used Markov Regime-Switching auto-regression model to capture the characteristics of offshore CNY exchange rate and its conversion mechanism
 - Applied EGARCH model to exploring the volatility of offshore CNY exchange rate
 - Compared the performance of the MS-EGARCH model with the single-state EGARCH model in the in-sample parameter estimation and the prediction of out-of-sample volatility using R

Data Science Camp, China Academy of Sciences

Beijing, China

- **An improved method for predicting the correlation between mRNA and human diseases** Feb, 2018
Advisor Dr. Xing Chen
 - Literature search for association study between mRNA and human health
 - Combining HGIMDA(Heterogeneous graph inference for miRNA-disease association) and NRWRH(Network-based Random Walk with Restart on the Heterogeneous network) to improve the prediction performance
 - Performed Random-Walk algorithm using Matlab

PUBLICATION

Byzantine-Tolerant Distributed Multiclass Sparse Discriminant Analysis (Submitted)
One-Round Communication Efficient Distributed M-Estimation (Submitted)

TECHNICAL SKILLS

- **Computer skills:** R, Matlab, LaTeX, SAS, VS code, Git