YAJIE DUAN

Piscataway, NJ | yd254@rutgers.edu | (848)391-4674 | https://yajie1020.github.io/yajieduan/

EDUCATION

Ph.D. Candidate, Statistics 09/2019-Present
Rutgers University – New Brunswick New Jersy, USA

- **Research Interests:** Bigdata Analytics, Multivariate Analysis, Data Visualization, Deep Learning, Clustering
- **Relevant Courses:** Theory of Probability, Theory of Statistics, Advanced Theory of Statistics I&II, Advanced Probability Theory I&II, Interpretation of Data I&II, Stochastic Processes, Statistical Inference, Multivariate Statistics, Regression Theory, Advanced Time Series Analysis, Life Data Analysis

Bachelor of Science, Statistics 09/2015-07/2019
Southern University of Science and Technology Shenzhen, China

- **Relevant Courses:** Discrete Mathematics, Application of Stochastic Processes, Introduction to Bigdata Science, Algorithm Design and Analysis, Machine Learning, Probability Theory, Time Series Analysis, Computational Statistics, Bayesian Statistics, Statistical Linear Models, Sampling Survey, Mathematical Statistics, Multivariate Statistical Analysis, Nonparametric Statistical Methods

Visiting Student, UBC's Vancouver Summer Program 07/2016-08/2016
The University of British Columbia Vancouver, Canada

- **Relevant Courses:** International Politics, International Trade and Financial Markets

RESEARCH EXPERIENCE

Projection Pursuit Indices and Data Visualization Methods for Big Data 10/2020-Present
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Proposed new Projection Pursuit (PP) indices that can be used for bigdata, using a data compression method called "data nuggets" that reduces a large dataset into a smaller collection of data nuggets that maintain the data structure
- Developed static and dynamic graphical tools using proposed PP indices; implemented guided tours to generate interactive and efficient visualization for bigdata and detect clusters, outliers and other nonlinear structures
- Built packages in R and Python to implement proposed data visualization method for big data

Generative Modeling of Protein Loop Backbones 12/2020-Present
Rutgers University – New Brunswick *Research Assistant to Prof. Sijian Wang*

- Built recurrent neural network (RNN) models to generate novel and realistic protein loop backbone structures based on a database of structurally homologous loops for HCV protease
- Built a bidirectional Long short-term memory (LSTM) model to generate protein loop backbone sequentially and a new sequence-to-sequence Variational Autoencoder (VAE) to generate novel protein loops with various lengths
- Implemented in PyTorch via Python and evaluate the viability and novelty of the generated protein loop structures

Patient-Centered Assessment of Risk of Stroke vs. Bleeding 04/2021-09/2021
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Developed an assessment system for the risks of stroke vs. bleeding taking patient's personal fears of outcomes into account, with a proposed novel two-stage Deming regression model
- Implemented an algorithm that produces a graph with two regions corresponding to whether the patient should take anticoagulants based on the predicted risks of stroke and bleeding and the patient's fears of bleeding
- Built a web application for physicians to help prescribe anticoagulants based on the proposed methodology

Novel Estimation Methods for Particle Count by Dilution Series Data 06/2021-Present
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Proposed novel estimation methods for particle count by dilution series data of a solution stock based on censored Binomial and Poisson distributions; conducted simulation studies that showed good performance of the proposed methodology to estimate the original concentration of particles
- Built a package in R and a web application to implement proposed method and perform an automatic particle assay

Undergraduate Research in Biostatistics 07/2018-09/2018
Collaborative Center for Statistics in Science, Yale University *Research Assistant to Prof. Heping Zhang*

- Built Bayesian models via the Metropolis-Hastings algorithm to estimate probability distributions of the chances of live birth, conception, and pregnancy
- Implemented Convolutional Neural Network (CNN) for 3-D brain-imaging data via R to locate sub-regions of the brain that are associated with clinical outcomes
- Created a web calculator, *Prediction Calculator for Pregnancy Outcomes - Yale C2S2*, as part of the paper *A personalized medicine approach to Ovulation Induction/Ovarian Stimulation: Development of a predictive model and online calculator from level-I evidence* (under review)

PUBLICATIONS

- **Y. Duan** and J. Cabrera. "*A New Projection Pursuit Index for Big Data*". To be submitted.
- D. Sargsyan, **Y. Duan,** J. Cabrera, C. Ananth, J. Kostis. WJ. Kostis. "*Patient-Centered Assessment of One Year Risk of Stroke vs. Bleeding*". Submitted to American Heart Association Scientific Sessions 2021.
- **Y. Duan**, J. Cabrera, D. Sargsyan, C. Ananth, J. Kostis and WJ. Kostis. "A two-stage Deming regression model with applications to multiple disease risk assessment". To be submitted.
- **Y. Duan**, J. Cabrera, D. Sargsyan, C. Lin. "Novel Estimation Methods for Particle Count by Dilution Series Data". To be submitted.
- **Y. Duan**, C. Lu, C. Thai, S. Wang, S. Khare. "*Generative Modeling of Loop Backbones for HCV protease using LSTM and sequence-to-sequence Variational Autoencoder*". In Progress.
- **Y. Duan**, X. Wei, D. Zhang and G. Tian. "*Hypothesis Testing for the Homogeneity of Two Zero-and-one-inflated Poisson Populations*". Submitted to Journal of Statistical Computation and Simulation.
- **Y. Duan** and G. Tian. "*Type II Shifted Multivariate Asymmetric Laplace Distribution based on Mixture of Normal Distribution*". Under revision.
- D. Amaratunga, J. Cabrera, **Y. Duan,** D. Ghosh, M. Katehakis, C. Lin, J. Wang, W. Wang, A. Yadav. "*Bootstrap-based confidence and prediction intervals for forecasting COVID cases and deaths*". Under revision.

PROJECTS

Cognitive Status Prediction during Preclinical Alzheimer's Disease Phases 07/2020-08/2020

Rutgers University – New Brunswick

- Analyzed a dataset of 1500 adults' cerebrospinal fluid (CSF) biomarkers during a preclinical Alzheimer's disease (AD) phase between 2001 and 2016
- Built an Ordinal Logistic Regression model to predict the course of cognitive status during preclinical AD phases for each participant given their characteristics and CSF biomarkers, via R
- Found that CSF biomarkers in preclinical AD can predict cognitive decline and the relationships depend on age, education and parent dementia

Effects of Gymnastics Activity on Early Adult Bone Development 11/2019-12/2019

Rutgers University – New Brunswick

- Analyzed a dataset of 42 girls' records about bone development from a longitudinal study (1997-present)
- Built generalized additive mixed models (GAMM) for the dataset via R; quantified the general effect of gymnastics on bone development and the contribution of early gymnastics on bone after one quits based on estimated parameters
- Found that the effects of gymnastics on bone development are different in vaiours body regions

SELECTED AWARDS AND HONORS

University Award for Outstanding Graduates (top 2% in 931 undergraduates) 2019
University Award for Outstanding Undergraduate Thesis (top 3% in 931 undergraduates) 2019
First Prize, 5th National Data Mining Competition in China (top 10 in 2542 teams) 2017
First-class University Scholarship for Outstanding Undergraduate (top 3% in 931 undergraduates) 2016

ACTIVITIES AND LEADERSHIP

Student Statistical Consultant 09/2021-Present

Office of Statistical Consulting, Rutgers University – New Brunswick

- Provide statistical advice for clients from both academic and industry
- Performed different statistical analysis models and provided guidance to non-statisticians across diverse disciplines.

Statistics Seminars Organizer 10/2017-05/2018

Southern University of Science and Technology

- Organized discussion sessions on computational tools in statistics (i.e., EM, MM, QLB algorithms)
- Offered classes with instruction on chapters in *Statistical Learning with Sparsity: The Lasso and the Generalization*

Vice President of Undergraduate Students' Union 09/2016-07/2017

Southern University of Science and Technology

- Led and managed the students' union, organized students' activities, and represented students to communicate with university

PROGRAMMING

R, Python, MATLAB, JAVA, LaTex, HTML, PHP, AJAX, JavaScript