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:** Biostatistics, Bigdata Analytics, Multivariate Analysis, Data Mining Methodology, Deep Learning
- **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, Bayesian Analysis

Bachelor of Science, Statistics 09/2015-07/2019

Southern University of Science and Technology Shenzhen, China

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

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

WORK EXPERIENCE

Summer Intern, Biostatistics 05/2022-08/2022

RWE Statistics Group, Pfizer Inc. New York, NY

- Derived theoretical estimates of predictions and marginal effects with standard errors for two-part model, a regression model for fitting zero-inflated outcomes; Implemented in R for the consistency with Stata package *twopm*
- Developed an S4-class R package *twopartm* that fits two-part models and provides predictions, average marginal effects and predictive margins with confidence intervals (*The R package is published on CRAN*)
- Literature review about impact of unmeasured confounders in observational studies including E-value method; Designed a simulation study comparing existing methods to quantify effects of unmeasured confounders

Research Assistant, Biostatistics 04/2021-09/2021

Cardiovascular Institute of New Jersey, Rutgers Medical School (directed by John B. Kostis, MD; closed in 2022) New Brunswick, NJ

- 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 proposed methods in R that produces a graph providing recommendations for patients about taking anticoagulants, based on the predicted risks of stroke and bleeding and the patient's fears of bleeding
- Built an R Shiny app for physicians to help prescribe anticoagulants based on the proposed methodology

Student Statistical Consultant 09/2021-01/2023

Office of Statistical Consulting, Rutgers University – New Brunswick Piscataway, NJ

- Performed statistical analysis such as hypothesis tests and regression models in pediatric orthopedics research studies at Rutgers medical school, including systematic scoping reviews on SPATT vs SPOTT in the treatment of cerebral palsy patients, shoulders, forearm, and elbow secondary surgical procedures in Neonatal Brachial Plexus Palsy, etc.
- Provided statistical advice and guidance to clients across diverse disciplines from both academic and industry

PUBLICATIONS

- Sargsyan, D., **Duan, Y.**, Kostis, W. J., Ananth, C., Kostis, J. B., Cabrera, J., & Myocardial Infarction Data Acquisition System (MIDAS) Study Group. (2021). Patient-centered Assessment of Risk of Stroke vs. Bleeding in Patients with Atrial Fibrillation. *Circulation, 144*(Suppl_1), A13362-A13362.
- Lin, C., **Duan, Y.**, Sargsyan, D., Cabrera, J., Livingston, C., Vogel, R., Hartman, J., Das, M., Talloen, W., Geys, H., Kanoulas, E. & Mohanty, S. Automated Spot Counting in Microbiology. *Accepted by IEEE/ACM Transactions on Computational Biology and Bioinformatics.*
- **Duan, Y.**, Cabrera, J., Sargsyan, D. & Lin, C. Particle Concentration Estimation in Dilution Series Experiments. *Accepted by Naval Research Logistics.*
- Beavers, T., Cheng, G., **Duan, Y.**, Cabrera, J., Lubomirski, M., Amaratunga, D. & Teigler, J. Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. *Submitted to Journal of Computational and Graphical Statistics.*
- **Duan, Y.** & Cabrera, J. A New Projection Pursuit Index for Big Data. *Submitted to Statistics and Computing.*

- **Duan, Y.**, Cabrera, J., Lin, C. & Sargsyan, D. A Bayesian Hierarchical Model for Particle Count Estimation in Dilution Series Experiments. *In Progress.*
- **Duan, Y.**, Cabrera, J., Sargsyan, D., Anath, C., Kostis, J.B. & Kostis, W.J. A Two-stage Deming Regression Model with Applications to Multiple Disease Risk Assessment. *Under revision.*
- Amaratunga, D., Cabrera, J., **Duan, Y.**, Ghosh, D., Katehakis, M., Lin, C., Wang, J., Wang, W. & Yadav, A. Adaptive learning models and techniques for forecasting COVID-19 daily cases. *Accepted by IEEE Transactions on Big Data.*
- **Duan, Y.**, Wang, J., Cabrera, J., Amaratunga, D., Katehakis, M. & Lin, C. COVID-19 Daily Case and Death Prediction using deep learning models with Time-lag Features. *Under revision.*
- **Duan, Y.**, Lu, C., Thai, C., Wang, G., Wang, S. & Khare, S. Generative Modeling of Loop Backbones for HCV protease using sequence-to-sequence Variational Autoencoder with attention mechanisms. *Under revision.*
- **Duan, Y.**, Cabrera, J. & Emir, B. twopartm: Two-Part Model with Marginal Effects in R. *Under review by The R journal.*
- **Duan, Y.**, Cabrera, J. & Gong, R. A Novel Error-in-variable Model Framework for Privacy-protected Data Analysis. *In Progress.*
- **Duan, Y.**, Wei, X., Zhang, D. & Tian, G. Hypothesis Testing for the Homogeneity of Two Zero-and-one-inflated Poisson Population. *Submitted to Journal of Statistical Computation and Simulation.*
- **Duan, Y.** & Tian, G. Type II Shifted Multivariate Asymmetric Laplace Distribution based on Mixture of Normal Distribution. *Under revision.*

PRESENTATIONS

- **Oral presentation:** "A New Projection Pursuit Index for Big Data." *INFORMS 2022 Annual Meeting* (October 2022, Indianapolis, IN, USA)
- **Oral presentation:** "COVID-19 Daily Case and Death Prediction using Deep Learning Models with Time-lag Features." *32nd European Conference on Operational Research (EURO 2022)* (July 2022, Espoo, Finland)
- **Poster presentation:** "Particle Count Estimation in Dilution Series Experiments." *2022 Conference on Advances in Bayesian and Frequentist Statistics* (April 2022, Piscataway, NJ, USA)

RESEARCH EXPERIENCE

Projection Pursuit Indices and Data Visualization Methods for Big Data 12/2020-Present
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Proposed new Projection Pursuit (PP) indices to find structures in big data, 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 to detect clusters, outliers and other nonlinear structures in bigdata; implementing guided tours to generate interactive and efficient visualization for big data
- Building packages in R to implement proposed data visualization method for big data; Developing differential PP indices to detect changes in distributions or clusters of big data

Novel Estimation Methodology for Particle Count in Dilution Series Experiments 09/2021-Present
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Proposed novel estimation methods for particle count in a solution by dilution series data from experiments, based on censored Binomial and Poisson distributions; conducted simulation studies that showed good performance of the proposed methodology to estimate the concentration of particles in neat samples
- Built a package in R to implement proposed method to perform an automatic particle assay with count estimates
- Developing Bayesian hierarchical models to estimate posterior distributions of particle counts; Developing a novel and efficient experimental design for serial dilution assays

Deep Learning Models for COVID-19 Case and Death Prediction 10/2021-06/2022
Rutgers University – New Brunswick *Research Assistant to Prof. Javier Cabrera*

- Developed a new methodology considering time lags to forecast both COVID-19 daily cases and deaths with confidence and prediction intervals, using Long short-term memory (LSTM) with proposed modified versions of variance stabilizing transformations and smoothing splines to deal with data irregularities and inconsistency
- Illustrated using New Jersey and Florida COVID-19 data with a good prediction performance

Generative Modeling of Protein Loop Backbones 10/2020-05/2022
Rutgers University – New Brunswick *Research Assistant to Prof. Sijian Wang*

- Developed a novel RNN (Recurrent Neural Network)-based sequence-to-sequence Variational Autoencoder (VAE) with attention mechanisms to generate novel and realistic protein loop backbone structures based on a database of structurally homologous loops for HCV protease
- Implemented via PyTorch in Python and evaluate the viability and novelty of the generated protein loop structures

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 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, Fertility and Sterility, 117(2), pp.408-418.*

SELECTED AWARDS AND HONORS

Travel/research funding by the School of Graduate Studies, Rutgers University	04/2022, 12/2022
University Award for Outstanding Undergraduate Thesis (top 1% in 931 undergraduates)	05/2019
First Prize, 5th National Data Mining Competition in China (top 10 in 2542 teams)	11/2017
First-class University Scholarship for Outstanding Undergraduate (top 2% in 931 undergraduates)	05/2016

ACTIVITIES AND LEADERSHIP

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

PROGRAMMING

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