Yu (Zoey) Zhu

Email : yzhu201@ucsc.edu
Linkedin: https://www.linkedin.com/in/yu-zoey-zhu-8a330113a/
Mobile : +1-(530)-304-6852
Personal Web: https://yuzoeyzhu.github.io/

Education

- **University of California, Santa Cruz** — Santa Cruz, CA
  PhD in Statistics; GPA: 4.0 — *Sep 2020 - Jun 2025*
- **University of California, Davis** — Davis, CA
  MS in Statistics; GPA: 3.88 — *Sep 2017 - Jun 2019*

Skills Summary

- **Languages**: R, Python, SQL, Java, C++, JavaScript
- **Tools**: GIT, Docker, Springboot, JIRA, Matlab
- **Data Manipulation**: Data Cleaning, Exploratory Data Analysis, Data Visualization
- **Hypothesis Testing**: A/B Testing, Test Design and Interpretation
- **Statistical Modeling**: Bayesian Parametric/Non-parametric Modeling, Time Series Analysis, Statistical Learning (GLMs, Tree-based Methods, SVMs, Unsupervised Learning, etc.), Model Selection, Feature Engineering, Causal Inference, Optimization

Honors And Rewards

2022 UCSC Statistics Summer Research Fellowship
2022 UW Biostatistics Summer Institutes Scholarship
2016 National Scholarship for Students with Excellent Academic Performance
2015 University Scholarship for Students with Excellent Academic Performance
2015 Outstanding Leader of Student Association for Science and Technology

Work Experience

- **ThoughtWorks** — Shanghai, China
  Software Development Engineer — *Aug 2019 - Aug 2020*
  - **Back-end SDE for Coca-Cola China Consumer Engagement Platform**: Developed back-end tasks, authorized winners and lottery campaign regulations. Broadcasted and implemented forwarding function back-end API design. Data migration and data source management. Maintained back-end unit test.
  - **Front-end SDE for IKEA PAX Cabinet AI Design System**: Participated in an recommendation and self-design system for customers to DIY, visualize and purchase their own PAX cabinet from scratch. Communicated UI designs and UI issues with client. Supported the production release.
  - **Front-end SDE for Starbucks APP**: Worked on the Starbucks application functions to help customers redeem their stars for coffee and other gifts. Helped scope features that attracted customers to use APP more actively.
- **Bank of Suning** — Nanjing, China
  Intern - Risk Compliance Department, Big-data Team — *Jun 2018 - Sep 2018*
  - Extracted, integrated, and cleaned up 3 million rows of customer transaction dataset; performed exploratory data analysis of patterns in missing data and feature distributions; detected outliers with Isolation Forest in Python.
  - Conducted SMOTE as the up-sampling method for the unbalanced data; applied Principle Component Analysis (PCA) to reduce data dimension; compared different machine learning models such as K-Nearest Neighbors, SVM with RBF Kernels and Ensemble Methods to predict the target customers for the financial loan product.
  - Selected Random Forest as the final model by AUROC comparisons on after hyper-parameter tunning under Grid Searching with 5-fold cross-validation.

Academic Research

Bayesian Hierarchical Model for COVID-19 Cause-of-death Prediction — Santa Cruz, CA
Research Assistant - Prof. Zehang (Richard) Li — *Sep 2020 - Current*

- Developed Bayesian hierarchical models with Latent structure to infer the prevalence of COVID-19 related death changing over time under the Verbal Autopsy data set in Brazil 2021.
- Applied structure priors for Multi-level Regression and Post-stratification (MRP) to better borrow information from neighbor cells when the sub-population cell becomes more refined and contains less information.
- Compared the independent structure as baseline form with Random Walk (RW) specification to validate the improvement of amount of information sharing between cells for the inference as well as the individual cause-of-death prediction.

Bayesian Non-parametric Bernstein Polynomial Model for ROC Curve Santa Cruz, CA
Research Assistant - Prof. Zehang (Richard) Li, Claudia Wehrhahn *Apr 2022 - Current*

- Proposed to model the Receiver Operating Characteristic (ROC) curve to validate a newly designed system for performing portable molecular diagnostic testing which denominated solid state PCR.
- Developed a flexible covariate dependent Bayesian non-parametric Bernstein polynomial model using stick-breaking process to accommodate to the bounded outcomes of the SS-PCR test.

Stochastic Nearest Neighbor Multiple Imputation of the TAST Database Davis, CA
Research Assistant - Prof. James Sharpnack *Feb 2019 - Dec 2019*

- Developed a new python package 'SDataFrame' to realize similar functions such as 'groupby' for data frame in Pandas, based on the imputation method of Stochastic Nearest Neighbors (SNN) with Euclidean distance.
- Simulated data with missing at random (MAR) missingness under Beta distribution with Guassian Mixture Model; Proposed SNN multiple imputation methods to compare with Multivariate Imputation by Chained Equations (MICE) as well as MissForest, and presented the advantages of SNN.
- Applied the SNN to the real-world demography data set TAST with records largely missing.

Course Projects

Semi-Supervised Learning (SSL) on Causal and Anti-Causal Structure Santa Cruz, CA
Advisor: Prof. Zehang (Richard) Li *Sep 2022 - Current*

- Implemented the semi-generative model and conditional self-learning algorithm under the real-world Verbal Autopsy data to validate the assumption that SSL works for Anti-Causal other than Causal relationship.

Bayesian Non-parametric Approaches for Stochastic Order in ROC Analysis Santa Cruz, CA
Advisor: Prof. Anathasis Kottas *March 2022 - June 2022*

- Applied the Bayesian non-parametric approaches Dirichlet process mixtures (DPM) and Mixtures of finite Polya tree (MPT) with stochastic order constraint to model the ROC curve with a meaningful value of AUROC that strictly larger than 0.5.

Image Recognition with Bayesian CNN for Simpsons Characters Santa Cruz, CA
Advisor: Prof. Juhee Lee *March 2022 - June 2022*

- Proposed and compared the Non-Bayesian Convolutional Neural Network (regularCNN) with Bayesian Convolutional Neural Network (BayesCNN) with Variational Inference based on the predictive performance for the image recognition task under Simpsons data set.
- Measure the uncertainty estimation in BayesCNN and interpreted the uncertainty based on the 95% credible intervals of posterior predicted class assignment probabilities for some of the test images.

The Statistics Behind Perceptual Decision Making Santa Cruz, CA
Advisor: Prof. Raquel Prado *Sep 2020 - Dec 2020*

- Proposed and implemented a randomized block model and two mixed effect logistic regression models to explore the relationship between response time, evidence strength level, accuracy, and confidence in a Two-Choice Decision (TCD) framework. The results supported the assumption that the decision formation is based on the accumulated evidence provided by sensory stimuli.

Robust PCA and Extreme Classification Davis, CA
Advisor: Prof. Cho-Jui Hsieh *Nov 2017 - Dec 2017*

- Applied the ADMM algorithm to solve a robust PCA problem under the non-convex condition and tested in MNIST
- Conducted the Conjugate Gradient Descent algorithm to solve a multi-label classification problem with an extremely large number of labels in MATLAB and R

Conference

Objective Bayes Santa Cruz, CA
A Bayesian Hierarchical Model for COVID-19 Related Cause-of-death Assignment Using Verbal Autopsies *Sep 2022*

Teaching Assistant

Fall 2021, Spring 2022 **[STAT 07] Statistical Methods for the Biological, Environmental and Health Science**
Fall 2020, WInter 2021, Winter 2022 **[STAT 05] Statistics**