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, Machine Learning Frameworks: Keras, PyTorch, Tensorflow
- **Tools**: GIT, Docker, Springboot, JIRA, Matlab
- **Data Manipulation**: Data Cleaning, Exploratory Data Analysis, Data Visualization
- **Experimentation**: A/B Testing, Experimental Design, Hypothesis Testing and Interpretation
- **Statistical Modeling**: Bayesian Hierarchical Modeling, Bayesian Parametric/Non-parametric Modeling, Causal Inference, Deep Learning (CNN, LSTM, GAN, etc.), Machine Learning (GLMs, Tree-based Methods, SVMs, Supervised/Semi-supervised/Unsupervised Learning, etc.), Reinforcement Learning, Data Mining, Time Series Analysis, Model Selection, Feature Engineering, Optimization, Distribution Shift, Domain Adaptation, Zero-shot Learning, LLM with prompt engineering and fine-tuning

Work Experience

- **Microsoft** — Redmond, WA
  Data Scientist Intern (Azure Core, Azure Compute team) — *June 2024 - Sept 2024*
  - **Reinforcement Learning for Azure Qualify A/B Testing Environmental Design**:
    - Developed an RL framework to optimize node configurations as environmental design for Azure pre-production testing.
    - Designed the RL model to dynamically adjust target distributions of hardware, virtue machine types and operating systems via incorporating the A/B testing statistical summary as feedback into the reward function.
    - Utilized Q-learning approaches to iteratively refine node configurations, ensuring improvements in signal validation.
  - **LLM with Root Cause Analysis**:
    - Developed a custom pipeline for Root Cause prediction and inference of regression incident from textual discussions, based on fine-tuning the Sequence for Classification model with LoRA adaptors and the Seq2Seq model which leverage the Chain-of-Thought (CoT) reasoning.
    - Designed and executed the efficient workflow combining the data pre-processing, tokenization and continuously re-training and re-validating the model.
- **Tencent America** — Palo Alto, CA
  Data Science Research Intern — *Oct 2023 - May 2024*
  - **Online Game Experimentation [CODE@MIT 2024]**:
    - Explored the challenge of A/B testing in online gaming, where dynamic and ephemeral **network interference** among players existed in user-randomized experiments, which compromised the validity of causal effects due to the SUTVA violations.
    - Proposed an innovative framework for **treatment effect estimation** tailored for scenarios where a completely randomized experimental design is implemented without explicit knowledge of network structures.
    - Developed the proposed method into an integrated pipeline for implementation within a real-game application.
  - **Media Channel Modeling for Game Advertising [Draft manuscript]**:
    - Developed a Bayesian hierarchical model framework to evaluate the effectiveness of different media channels for a PC game's online ads campaigns, showing a great fit with an **extremely short-term** advertising data set.
    - Incorporated the **Carryover and Shape Effects** to correlate advertising costs with installs, validated on the methodologies of the mixture of weighted functions from Jin et al., 2017.
    - Assessed the impact of various priors and effect functions on model performance with limited data. Proposed the **optimized advertising budget allocation** plan for the incoming ads campaign.
- **ThoughtWorks** — Shanghai, China
  Software Development Engineer (Full Time, Data Track) — *Aug 2019 - Aug 2020*
  - **Coca-Cola China Consumer Engagement Platform**: Designed and developed data processing pipelines using Python to extract, transform, and load data from various sources. Developed algorithms to analyze consumer engagement data and generate insights for business decision-making.
  - **IKEA PAX Cabinet AI Design System**: Developed machine learning algorithms to recommend personalized design options for customers based on their preferences and past behavior. Optimized the system's performance by analyzing user data and implementing feature engineering techniques.
  - **Starbucks APP**: Analyzed user engagement data using SQL and Python to identify patterns and trends in customer behavior. Developed machine learning models to predict user churn and optimize the customer reward program.

Academic Research

Bayesian Latent Variable Models for Mortality Surveillance Santa Cruz, CA
Research Assistant - Prof. Zehang (Richard) Li *Sep 2020 - Current*

- Developed **Bayesian hierarchical models with latent structure** to infer the stratum-specific prevalence (SSP) of COVID-19 related death stratified with time and age with **distribution shift** across domains.
- Demonstrated the formal framework to analyze **partially verified** Verbal Autopsy data under the **non-ignorable training data selection mechanism.**
- Introduced the novel use of **structured priors** to improve prevalence estimation for small sub-populations by more efficiently borrowing information from different sub-populations.
- Improved flexibility in modeling symptom distributions with respect to cause of death by incorporating advanced **tensor decomposition** techniques to capture the clusters of symptoms and the corresponding dependence.

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.

Course Projects

Semi-Supervised Learning (SSL) on Causal and Anti-Causal Structure [Report] 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 [Report] 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 [Report] 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.

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

CODE@MIT Boston, MA
Treatment Effect Estimation Amidst Dynamic Network Interference in Online Gaming Experiments [Talk] *Oct 2024*
JSM Portland, OR
A Bayesian Hierarchical Model for Mortality Surveillance using Verbal Autopsy [Talk] *Aug 2024*
WNAR Anchorage, AK
A Bayesian Hierarchical Model for Mortality Surveillance using Partially Verified Verbal Autopsy Data [Talk] *Jun 2023*
Objective Bayes Santa Cruz, CA
A Bayesian Hierarchical Model for COVID-19 Related Cause-of-death Assignment Using Verbal Autopsies [Poster] *Sep 2022*

Honors And Rewards

2024 SFASA Student Travel Award
2024 UCSC Graduate Dean's Travel Fellowship
2023 UCSC Statistics Summer Research Fellowship
2023 UCSC Graduate Dean's Travel Fellowship
2023 WNAR Student Paper Competition Travel Fellowship
2022 UCSC Statistics Summer Research Fellowship
2022 UW Biostatistics Summer Institutes Scholarship
2016 National Scholarship for Students with Excellent Academic Performance

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**