Haoxuan Wang

Contact Information: whxsavitar@gmail.com | (919) 433-7125 | master-savitar.github.io | github.com/Master-Savitar

Research Interests: Bayesian hierarchical modeling, Gaussian processes, Density estimation, Statistical methods with application to the biological community and public health

Education

Duke University Durham, NC, USA
Master of Science in Statistical Science Aug 2022 – May 2024 (Expected)

- Cumulative GPA 3.957/4.00
- Courses: Bayesian Statistical Modeling, Statistical Inference, High-dimensional Statistics, Real Analysis, etc.

Xi'an Jiaotong University Xi'an, China
B.Mgt. in Big Data Management and Application Aug 2018 – July 2022

- Cumulative Average 91.21% (GPA 3.93/4.00, Rank 1/28)
- Courses: Probability Thoery & Mathematical Statistics, Machine Learning, Optimization, Numerical Analysis, etc.

Working Papers & Publications

- **Wang, H.**, Lauha, P.M., Dunson, D.B. (2023). Bayesian Modeling of Multi-species Labeling Errors in Ecological Studies, to be submitted to the *Journal of the American Statistical Association.*

Research Experience

Data Augmentation for a Gaussian Process Density Model Jan 2023 – Present
Advisor: Surya Tapas Tokdar, Professor of Statistical Science, Duke University

- Proposed a novel nonparametric density model based on a Gaussian Process (GP) prior, where the data generation process is characterized by a Perfect Binary Tree (PBT), incorporating comparisons and rejections of latent proposals
- Established the weak posterior consistency of the PBT-GP prior at continuous true densities when employing a common covariance kernel and specifying an appropriate prior on the covariance kernel's smoothing parameter
- Obtained a joint likelihood that conjugates to the GP prior by augmenting the model with the rejection history of the data generation process and Pólya-Gamma random variables, allowing for efficient inference with Gibbs sampling
- Optimized computational performance by approximating the smooth GP with a low-rank predictive process; demonstrated our approach's superiority in density estimation and the mixing time via extensive simulation studies

Bayesian Hierarchical Modeling of Multi-species Labeling Errors Oct 2022 – Present
Advisor: David B. Dunson, Arts and Sciences Distinguished Professor of Statistical Science, Duke University

- Developed a Bayesian hierarchical modeling framework tailored to the multi-label crowdsourcing tasks, which considers correlations among bird species, models species occurrences using a Dirichlet process (DP) Bernoulli Mixture Model and accommodates the variability in the quality of bird experts annotations across different bird species
- Employed informative and conjugate priors for model parameters to address challenges posed by the sparsity of the species annotation data; applied Pólya-Gamma augmentation for efficient inference through collapsed Gibbs samplers
- Demonstrated the effectiveness of our framework in aggregating annotations and estimating experts' skills in terms of the coverage of 95% posterior credible interval and mean squared error (MSE) through numerical experiments; applied this framework to a dataset of Finnish bird vocalizations annotated by experts in a crowdsourcing project

Tensor Completion with Side Information & its Application in Recommender Systems Aug 2021 – Aug 2022
Advisor: Yao Wang, Professor of School of Management, Xi'an Jiaotong University

- Established a mathematical representation of multiple graphs with shared vertices; introduced a graph smoothness regularization technique for multiple graphs that is specifically adapted to three-order tensor completion
- Integrated transformed tensor-singular Value Decomposition (t-SVD) with our graph smoothness regularization; developed an effective optimization algorithm leveraging the alternating direction method of multiplier (ADMM)
- Conducted comprehensive numerical experiments on both synthetic and several public datasets for recommender systems, demonstrating the effectiveness of the proposed method in scenarios with limited data availability

Multi-dimensional Evaluation System Based on Hotel Reviews Aug 2020 – May 2021

Advisor: Shaolong Sun, Distinguished Fellow of School of Management, Xi'an Jiaotong University

- Compiled more than 100k tourists' reviews with Python crawler; trained Doc2Vec (an adaptation of Word2Vec) with the preprocessed reviews and performed sentiment analysis using the k-means algorithm
- Applied Latent Dirichlet Allocation (LDA) with Gibbs sampling to identify underlying topics, incorporating the perplexity as a metric to choose the number of topics; extracted five key topics pertaining to the hotel situation
- Determined each review's topic composition with the trained LDA model; computed sentiment scores for reviews on five different topics and proposed targeted recommendations beneficial for both hotels and tourists

Professional Experience

Institute for Interdisciplinary Information Core Technology Xi'an, China

Intern, Quantitative Researcher Oct 2021 – Mar 2022

- Constructed an active manager factor and a few fundamental factors; evaluated their stability and consistency of the impact on stock returns with Information Coefficients (IC) and demonstrated the effectiveness with portfolio sort test
- Trained many-to-many RNNs with LSTM, GRU layers separately using high-frequency trading data, optimizing with negative IC as the loss function and employing early stopping, dropout to alleviate overfitting; demonstrated the factor obtained by GRU is more effective in terms of accuracy

COMPETITION

IKCEST "Belt and Road" International Big Data Competition May 2020 – Sep 2020

Excellence Award, Rank 54/3023

- Constructed a modified SEIR model incorporating the effect of both intercity migration and population flow between regions within every city to predict the epidemic trend of the COVID-19 pandemic in five virtual cities
- Proposed a method of estimating and predicting the dynamic parameters in the modified SEIR models for five virtual cities and every key region separately with the LSTM model according to the historical number of newly infected persons, intercity migration scale index as well as grid population flow index
- Trained our models with labeled data over the first sixty days and accurately predicted the number of newly infected persons per day over the last thirty days with a root mean squared logarithmic error (RMSLE) of 1.42427

Awards & Membership

- Dean's Research Award for Master's Students, Duke University Nov 2023
- Eastern North American Region (ENAR) Membership since Sep 2023
- Outstanding Graduate Award, Xi'an Jiaotong University Jun 2022
- Chen & Zhu Economics and Management Scholarship, Xi'an Jiaotong University Jan 2021
- Full Scholarship Exchange to The University of Hong Kong Oct 2020
- Excellence in the 2nd IKCEST "Belt and Road" International Big Data Competition **(54/3023)** Sep 2020
- Honorable Mention of the Mathematical Contest in Modeling (MCM) Apr 2020
- The First Prize of the Shaanxi Province in National College Students Mathematical Contest **(5%)** Oct 2019
- The First Class Scholarship, Xi'an Jiaotong University **(1%)** Sep 2019
- The First Prize in the 34th High School Chinese Physics Olympiad **(Top 50 Provincial Ranking)** Oct 2017

Skills

Programming:

- R Environment: RStan, RCpp, Tidyverse, ggplot2, Web Scraping, Packages Development, Text Mining, Shiny App
- Python: NumPy, SciPy, Pandas, Scikit-learn, Seaborn, PyTorch, Tensorflow, Scrapy
- Others: C/C++, Java, MATLAB, MySQL, LaTeX, Git

Languages: Mandarin Chinese (Native) and English (Proficient)

Personal Interests: Passionate about Soft Brush Calligraphy (Certified Level 10), Watercolor Painting, reading books, playing basketball, hiking and volunteer work