AIHUA LI

Email: aihua.li@duke.edu

Research paper and other published work are accessible via hyperlinks highlighted in blue.

EDUCATION

Shanghai University of Finance and Economics, 2016 - 2020
Bachelor, Department of Statistics and Management

- Major: Statistics, GPA: 3.75/4.0, rank: 1/78
- Core courses: Advanced Algebra, Mathematical Analysis, Probability theory, Regression, Numerical Analysis, Multivariate Statistics, Stochastic Processes

Duke University, 2020 - 2022
Master's in Statistical Science, Department of Statistical Science

- Major: Statistics, GPA: 3.96/4.0, rank: Top 1% in 40
- Core courses: Predictive Modeling, Bayesian Statistical Modeling, Theory of Inference, Hierarchical Models, Categorical Data, High Dimensional Data

RESEARCH EXPERIENCES

Pitfalls of Bayesian spike and slab priors and a decoupled shrinkage and selection approach to sparse estimation
July 2021 - In progress *First Advisor: Surya Tokdar, Secondary Advisor: Jason Xu*

- The spike and slab prior, a gold standard for Bayesian variable selection, has poor sparsity behaviors on high dimensional dependent data. In high dimensional problems, Bayesian's intrinsic protect against model complexity imposes a favor for parsimonious models, and may lead to an over-conservative selection. Furthermore, when there is data dependency, marginal inclusion probabilities are scattered across correlated predictors, which often fail to determine an effective median probability model. Motivated by the pitfalls of spike and slab priors, we introduced the idea of decoupled shrinkage and selection for sparse estimation.
- In Bayesian decoupling, sparse estimation was realized via Bayesian decision procedures by loss functions penalizing model complexity. Both L_0 and L_1 penalties were considered. Our simulations showed that in high dimensional dependent cases, the decoupled estimates provided an intermediate alternative to the over-shrunk median probability models or the large full models. Along the sparsification path, the contribution of predictors in data variation was decaying, and we could identify few representative predictors explaining almost all variation.
- Showing how the decoupled sparsification gives a well-interpretable variable selection result, we will then focus on the control over false discovery rates via Bayesian decoupling in our future work.

A generalized linear mixed-effects model for computational deconvolution on transcriptomics data
July 2021 - In progress *Advisor: Jichun Xie*

- Computational deconvolution is a developing technique to infer cell type proportions from bulk transcriptomics data. We proposed a deconvolution model with a focus on the heterogeneity in cell proportions among different patient subgroups. Considering the within-group correlation in gene expressions, we proposed both generalized estimating equations (GEE) and generalized linear mixed-effects models (GLMM).
- While simplified model-based assumptions are generally accepted in the applied field for feasibility and interpretability, we particularly focused on the model fitting and the validity of the statistical assumptions. In data aggregation, which was a necessary procedure for constructing deconvolution datasets, we emphasized the issues of mismatched data resolutions. Besides, the identifiability problem in mixed-effects models was discussed. Furthermore, we imposed constraints on coefficient estimates and considered constraint optimization.
- Through applications on real datasets for Glioblastoma brain cancer, we have shown the validity of the GEE and the GLMM as deconvolution methods. Future work will be focused on the problems of weak signals and possible Bayesian modeling to allow more flexible distributional assumptions.

Undergraduate Thesis: A Model Averaging Approach for Functional Linear Regression Models
January 2020 - May 2020

- In functional linear regression models, the orthogonal series estimation, a commonly used nonparametric approach to function estimation, necessitates a model selection procedure to decide on the order of the basis expansions. Motivated by the fact that model selection ignores the model uncertainty and often results in estimates with high sensitivity to data perturbation, we introduced the idea of model averaging as an alternative option to model selection, and proposed a model averaging type estimate based on Fourier orthogonal series estimation in functional linear regression models.
- In a simulation comparing the model averaging estimates with the estimates determined by the model selection criteria AIC and GCV, we showed the superiority of the former one in terms of the prediction performance measured by MSE. Besides, a real data application showed that the model averaging approach could substantially reduce the sensitivity to the sample data.

Conditional Estimates of Diffusion Processes for Evaluating the Positive Feedback Trading
January 2019 - May 2019 *Published paper on arXiv*

- Motivated by the relationship between positive feedback trading and investors cognitive bias, this research proposed a quantitative measurement of the bias based on the conditional estimates of diffusion processes. We proved the asymptotic properties of the proposed estimates. The asymptotic properties helped to interpret the investment behaviors that if a feedback trader finds a security perform better than his expectation, he will expect the future return to be higher, while in the long term, this bias will converge to zero.
- Furthermore, considering adaptive expectations in reality, we introduced an exponential smoothing method to adjust the return expectation closer to its fundamental level. An empirical study was done on 10 stocks from CSI 300 Index in China covering 10 years from 2009 to 2018, where we showed the effectiveness of the exponential smoothing adjustment which lowered the return forecast bias by 51.78%.

An Application of Reinforcement Learning in Trade-off between Safety and Traffic Congestion
June 2018 - March 2019 *Unpublished complete paper*

- This applied research was motivated by the alarming traffic accident rates caused by the poor traffic design at a university campus gate. The traffic design was not corrected due to traffic congestion.
- Based on reinforcement learning, we proposed a multi-agent system for controlling the traffic lights at multiple intersections. The system adaptively accounted for the waiting cars. The program was realized by the Q-learning algorithm. Showing the efficiency of the model in alleviating traffic congestion, we generalized its application for all similar situations and suggested the government improve the current inefficient traffic system.
- In September 2021, this project was referred by the school committee and the local government. A long-pending proposal to build a traffic light at the school gate for students' safety was approved.

PROJECTS IN STATISTICS AND MATHEMATICS

Model Diagnostics Tools in Hierarchical Models and Comments
April 2021 *Published work online: Diagnostics-in-LMM*

- This was an independent study project in hierarchical models. We focused on the diagnostics procedures in linear mixed-effects models (LMM), and provided a comprehensive discussion on the difficulties in LMM diagnostics, including the issues of unstructured dependency assumptions, small groups, multilevel variables, interpretations of the influential observations, etc.
- Due to the high flexibility and complexity of LMMs, the diagnostics tools have been sloppily documented for practitioners. Therefore, we collected in one place the various tools for diagnostics along with the analysis, and made it public online via the above link.

Continuation Ratio Models on Ordinal Data with Smoothed Estimates of Time-varying Coefficients
October 2021

- This independent project explored a clinical dataset with over 2000 clinical records and 23 covariates, where the target response was the gestational age. A continuation ratio probit model was constructed on the gestational age as ordered categorical data. The goal was to infer the associations between the risk of premature delivery and the level of DDT exposure and PCB exposure, while adjusting for the impacts of other covariates.
- For the sparse observations problem in the extreme categories on the distribution tails, we introduced smoothing techniques by B-spline basis expansions for estimating the time-varying coefficients.
- For the correlated clinical measuresments, we did principal component analysis to construct representative variables. Besides, model selection was done by likelihood ratio tests among nested models.

Gaussian Mixture Models in Differential Gene Expression Analysis on Gastric Cancer Data
October 2020 - November 2020

- This independent project was a differential gene expression (DGE) analysis focusing on the associations between the gastric cancer stage and the expressions of 85 genes on the TDF β signalling pathway. A Gaussian mixture model was built on the normalized gene expression levels, and was realized by Bayesian MCMC sampling.
- Before modeling, data quality control included principal component analysis and hierarchical clustering to check the sample dependencies, grouping effects, and outliers. After modeling, posterior predictive checks were done to ensure the model fitting on the bimodal response.

PROJECTS IN COMPUTING

A Python Implementation of Biclustering by Sparse Singular Value Decomposition
April 2021 *Published package online: Biclustering-SSVD*

- This project wrote and published a packge in Python to implement the sparse singular value decomposition (SSVD) algorithm for biclustering proposed in the literature. Code optimization was realized via C-extensions in Python. We wrote a package document to describe the algorithm and make comparative analysis in computing accuracy and speed with competing algorithms, including singular value decomposition and sparse principal component analysis. See the package instruction and document via the above link.

An Interactive Shiny App for Heart Disease Prediction by Logistic Models
November 2020 *Published work online: Heart-Disease-Prediction*

- This was an R project for Shiny App. We wrote an interactive interface to visualize the statistics of heart diseases in the U.S. and provided a prediction machine by a logistic model. That prediction interface allowed users to evaluate their risks for heart disease based on their health statistics. The site is made public via the above link.

WORK EXPERIENCES

Teaching Assistant for Bayesian Statistics (Course Code: STA602, Duke University)
August 2021 - December 2021

- Taught and graded weekly in-person labs. Assisted in writing exercises and solutions.

HONORS

Dean's Research Award for Master's Students	2021
Outstanding Undergraduate Awards	2020
First Place in The People's Scholarship in China	2017, 2018, 2019
Tailong Bank Scholarship	2019
Outstanding in National Students' Innovation and Entrepreneurship Program	2019
National Scholarship	2018
Second Place in China Undergraduate Mathematical Contest in Modeling	2018
China Merchants Bank Scholarship	2017