Student Contributed Posters

1. Iterative Multivariate Random Forest for Feature Selection in Integrating Multi-Omics Datasets

Wei Zhang[1] and X. Steven Chen[1]
[1]Division of Biostatistics, Department of Public Health Sciences, University of Miami, Miller School of Medicine

Integrating multi-omics datasets requires the selection of important features from one data type that are associated with another data type. Traditional methods such as sparse canonical correlation analysis (sCCA) and sparse partial least squares (sPLS) face challenges in handling non-linear, mixed types of information, and interactions present in omics data. Tree-based methods such as random forest are more effective in handling such data. In this work, we propose an iterative multivariate random forest (iterMRF) approach to extract major contributing features between omics datasets. The method uses an iterative framework to select important features in the random forest with a multivariate splitting rule. The feature selection procedures were evaluated by several criteria, including permutation importance measures and variable splitting weights. Our approach shows competitive performance in selecting important features between data types through both simulation studies and real data cases.

2. Future of Data Science Emerging Technology Trends: A perspective analytics on large dataset of research databases

Muhammad Usman Aslam
Department of Science and Engineering (Data Science), University of West Florida

This undertaking analyses the evolution of Emerging Technologies around the globe, using the predictions made in the Horizon Report, published yearly from 2004, MIT Technology review and predictions made by Institute of Electrical and Electronics Engineers (IEEE). This research applies social evaluation, primarily based on Google Trends, and Bibliometric analysis, with data of scientific publications from IEEE, MIT, Questia online Library, JSTOR, Springer, Hindawi and Web of Science, with a purpose to discover which technology had been a hit and sincerely impacted mainstream training, and which one failed to have the anticipated impact. This mission gives guidelines that can be beneficial to those who are
looking forward to investing in new research regions.

3. Data-driven Evaluation of Trajectories in Single-Cell RNA-Seq Data

Xiaoru Dong[1] and Rhonda Bacher[1]

[1]Department of Biostatistics, University of Florida

Trajectory inference (TI) is an important component in single cell RNA-sequencing data, which orders single cells along trajectories based on the similarity of expression and consequently to model the process of dynamic cellular changes. However, a great number of preliminary tasks such as feature selection and dimension reduction need to be completed before proceeding to TI. We found that the choice of parameters in these preprocessing steps can greatly affect the reliability of the trajectory estimation. We created a novel 2-phase framework to evaluate what parameters should be used in preprocessing stage with the aim of obtaining more accurate trajectories in the TI analysis. Our 2-phase framework consists of detecting the existence of trajectories and evaluating the goodness of trajectory fitting, which assess credibility of the estimated trajectory in the high and low dimensional spaces based on the similarity between cells respectively.

4. Novel Spatial Two-Fold Modeling of Perimetric Rates of Change Using Glaucomatous Population Data

Chen Zhao[1]*, J. Sunil Rao[1], Swarup S. Swaminathan[2]
[1]Public Health Science, University of Miami
[2]University of Miami Health System, Bascom Palmer Eye Institute

Static automated perimetry is the most common form of measuring visual fields. In glaucomatous eyes, it is generally assumed that visual loss can be due to either a global loss or focal loss of visual sensitivities. In this work, we will consider modeling focal visual loss over time in a series of 500 glaucoma patients at the Bascom Palmer Eye Institute. Specifically, we have repeated visual field tests for each individual along with information on three different types of focal clustering patterns. The goal is to better identify fast progressors amongst this group of patients. To model this data, we develop two-fold random slope models which can be seen as an extension of the two-fold small area estimation models of Torabi and Rao (2014). We also include spatial structure in the fixed effects to account for the orientation of focal clusters. Our model borrows strength across individuals and across connected focal clusters resulting in more accurate estimates (as measured by estimated MSPE) of individual rates of visual sensitivity decline. Confidence intervals are estimated using parametric bootstrapping.

5. Neural-network transformation models for counting processes

Rongzi Liu[1], Chenxi Li[2], Qing Lu[1]

[1]Department of Biostatistics, University of Florida
[2]Michigan State University

While many survival models have been invented, the Cox model and the proportional odds model are among the most popular ones. Both models are special cases of the linear

transformation model. The linear transformation model typically assumes a linear function on covariates, which may not reflect the complex relationship between covariates and survival outcomes. Nonlinear functional form can also be specified in the linear transformation model. Nonetheless, the underlying functional form is unknown and mis-specifying it leads to biased estimates and reduced prediction accuracy of the model. To address this issue, we develop a neural-network transformation model. Similar to neural networks, the neural-network transformation model uses its hierarchical structure to learn complex features from simpler ones and is capable of approximating the underlying functional form of covariates. It also inherits advantages from the linear transformation model, making it applicable to both time-to-event analyses and recurrent event analyses. Simulations demonstrate that the neural-network transformation model outperforms the linear transformation model in terms of estimation and prediction accuracy when the covariate effects are nonlinear. The advantage of the new model over the linear transformation model is also illustrated via two real applications.

6. Assessing the impact of key parameters on the results of meta-analyses

Wenshan Han, Department of Statistics, Florida State University

Systematic reviews and meta-analyses are essential tools in contemporary evidence-based medicine. They have been extensively used to synthesize diverse sources of evidence and draw a conclusive and unified decision for medical decision-making. However, in recent years, conflicting conclusions from different meta-analyses conducted by separate teams on the same research topic have raised concerns about the reliability of these studies. One reason for this issue is the sensitivity of meta-analysis results to certain factors, such as the inclusion criteria for eligible studies in the systematic review and the choice of meta-analysis models. This essay aims to evaluate the conclusion of a meta-analysis using the arm-based model under the Bayesian framework. The arm-based meta-analysis model focuses on estimating the marginal results of treatment arms and provides flexibility for drawing conclusions with different effect measures. Despite its benefits, the arm-based model heavily relies on the correlation between treatment groups, a crucial parameter that accounts for randomization in clinical trials. Moreover, the between-study variance parameters play a critical role in meta-estimates by reflecting heterogeneity. Our proposed methods employ the spirit of tipping point analysis, a widely adopted technique in missing data imputation, to evaluate the robustness of meta-analysis results concerning the estimation of key parameters. Additionally, we introduce innovative visualization tools that intuitively display the impact of key parameters on meta-results. We apply these methods to two real-world meta-analyses and observe gradual changes in meta-analysis results as we assign key parameters to a series of values.

7. Single Cell Linear Adaptive Negative-binomial Expression Testing (scLANE)

Jack R. Leary[1] and Rhonda Bacher[1]
[1]Department of Biostatistics, University of Florida

Single cell RNA-sequencing (scRNA-seq) offers a high-resolution view of cellular biology, including dynamic processes such as differentiation and disease progression. Many methods have emerged that estimate a cell-level time ordering from static scRNA-seq samples, which use similarity of gene expression to place cells in order on some biological manifold. Researchers then typically 1) assume the ordering represents a biological process and 2) characterize which genes are associated with that process. Changes in expression over trajectories are usually complex with generalized additive models (GAMs) the dominant current choice of model. However, while GAMs are excellent for fitting nonlinear relationships, they are not easily interpretable. To address this trade-off, we developed Single Cell Linear Adaptive Negative-binomial Expression (scLANE) testing. scLANE balances the need for a flexible nonlinear model and the facilitation of biological interpretation. We demonstrate our method,Äôs accuracy and ability to draw meaningful comparisons from the model on simulated data and a case-study datasets having tens of thousands of cells and from multiple subjects.

8. Cherry Blossom Prediction

Nitul Singha, Department of Mathematics & Statistics, University Of West Florida

The blooming of cherry blossoms is a highly anticipated event in many parts of the world, attracting tourists and locals alike to witness the beauty of these delicate flowers. However, predicting the arrival of cherry blossoms has become a complex and fascinating challenge, as weather patterns and climate change can greatly impact the timing of their blooms. In 2023, the American Statistical Association (ASA) is hosting a cherry blossom prediction competition, inviting students, researchers, and citizen scientists to predict the peak bloom date of cherry trees in four locations around the world: Washington D.C., USA; Kyoto, Japan; Vancouver, Canada; and Liestal-Weideli, Switzerland.

The competition presents a unique opportunity for participants to apply their skills in predictive modeling to a real-world problem, as predicting cherry blossom blooms has important implications for tourism, agriculture, and ecology. The data available for the competition dates back to 1981, allowing participants to explore patterns in cherry blossom phenology over the past few decades. However, accurately predicting the timing of cherry blossom blooms remains a significant challenge due to the complexity of weather patterns and the impact of climate change on plant phenology.

To address this challenge, we will be using two common predictive modeling techniques in our project: logistic regression and linear regression. Logistic regression is a widely used statistical technique that is particularly effective for binary outcomes, such as the ""month"" of cherry blossoms. Linear regression, on the other hand, is a powerful tool for predicting numerical outcomes, such as the timing of cherry blossom blooms. By analyzing historical data from the four target cities and applying these two techniques, we hope to develop accurate models that can make reliable predictions for the 2023 competition.

Our project will contribute to the scientific understanding of cherry blossom phenology and the impact of climate change on plant growth and development. The findings of our study may also have practical applications in fields such as agriculture and tourism, where accurate predictions of cherry blossom blooms can have important economic and cultural implications. We believe that our project will help further the development of predictive modeling techniques and provide valuable insights for future prediction competitions.

9. A New Regularized Algorithm for Variable Selection using Hidden Markov Model in Longitudinal Data Analysis

Man Chong (Henry) Leong, Department of Biostatistics, University of Florida

Introduction: Modern longitudinal data analysis (LDA) often involves high-dimensional data with temporal dependence and mixtures, resulting in repeated measurements for each subject that are neither independent nor from a single distribution. Hidden Markov Models (HMMs) are popular tools in LDA, but their capacity is limited in high-dimensional settings due to a lack of dimension reduction or variable selection techniques. To address these challenges, we aim to develop a novel algorithm for using HMMs in high-dimensional LDA by combining HMMs and regularized generalized linear models (GLMs). Methods: Our proposed algorithm involves two steps: variable selection and maximum likelihood (ML) estimation. In the variable selection step, we use a combination of the coordinate descent algorithm and the EM algorithm to select a subset of significant covariates, with hyperparameters determined through cross-validation. The selected covariates from the first step are then used to refit an HMM using unregularized ML estimation. The performance of the algorithms is evaluated through simulation studies. Results: The simulation results demonstrate that our method can accurately select the set of covariates with non-zero underlying values and provide unbiased estimators for both the Markov chain and the selected regression model parameters. In comparison to using regularized GLMs for mixture data in LDA, our method shows improved performance, with a false negative rate up to 6.5% lower in variable selection. Conclusion: We present a novel two-step algorithm for modeling high-dimensional data with temporal dependence and mixtures in LDA using regularized HMMs. The proposed algorithm performs variable selection in the first step and provides unbiased estimates with ML in the second step. We show that the algorithm can identify the significant regression model parameters and give unbiased estimates to both the Markov chain and the selected regression model parameters through simulations."

10. Pseudo-Value Regression of Clustered Current Status Data with Informative Cluster or Subcluster Sizes in a Multistate Model

Samuel Anyaso-Samuel[1*], Dipankar Bandyopadhyay[2], Somnath Datta[1]
[1]Department of Biostatistics, University of Florida
[2]Virginia Commonwealth University

Current status data presents a more severe form of censoring due to the single observation of study units transitioning through a sequence of well-defined disease states at random inspection times. The current status data may be clustered within specified groups, and informativeness of the cluster sizes may arise due to a relationship between the transition outcomes and the cluster sizes. Failure to adjust for this informativeness will lead to a biased inference. Motivated by estimating covariate effects on the state occupation probability (SOP), we propose using the pseudo-value approach to model clustered current-status data with informative cluster or subcluster sizes. First, we formulate the pseudo-values by marginal estimators of the SOP computed using nonparametric regression theory. Secondly, the estimating equations based on the pseudo-values are reweighted by functions of the cluster sizes to adjust for informativeness. We perform simulation studies to investigate the pseudo-value regression under different scenarios of informativeness. The method is applied to a motivating periodontal disease dataset which encapsulates the complex data-generating mechanism.