Student Contributed Presentations

Contributed Session A

1. S-GMAS: Shape based Genome-wide Mediation Analysis

Shengxian Ding[1*], Rongjie Liu[1], Anuj Srivastava[1] and Chao Huang[1]
[1]Department of Statistics, Florida State University

Causal mediation analysis is widely utilized in neuroscience to investigate the role of brain image phenotypes in the neurological pathways from genetic exposures to clinical outcomes. However, it is still difficult to conduct a genome-wide mediation analysis with the shapes of brain regions as mediators due to several challenges including (i) large-scale genetic exposures, i.e., millions of single-nucleotide polymorphisms (SNPs); (ii) nonlinear Hilbert space for shape mediators; and (iii) statistical inference on the direct and indirect effects. To tackle these challenges, this paper proposes a mediation analysis framework with high dimensional genetic exposures and shape mediators. First, the square-root velocity function representations are extracted from the shapes, which fall in an unconstrained linear Hilbert subspace. Second, to address the issue caused by the high dimensionality in genetic exposures, the global sure independence screening procedure is conducted to discover candidate SNPs influencing the shape mediators. To identify the underlying causal pathways from the detected SNPs to the clinical outcome implicitly through the shape mediators, we proposed a framework consisting of a shape-on-scalar model and a scalar-on-shape model. Furthermore, the bootstrap resampling approach is adopted to investigate both global and local significant mediation effects. Finally, our framework is applied to the ADNI corpus callosum shape data and we successfully identify the mediation effect of a subset of candidate SNPs on Alzheimer's Disease through a subregion of the corpus callosum.

2. Multimodal Functional Deep Learning for Multi-omics Data

Yuan Zhou[1*], Dr. Shan Zhang[2], Dr. Pei Gang[3], Dr. Qing Lu[1]
[1]Department of Biostatistics, University of Florida
[2]Michigan State University
[3]Illinois State University

With rapidly evolving high-throughput technologies and ever-decreasing costs, it becomes feasible to collect diverse types of omics data in large-scale studies. While the multi-omics data generated from these studies hold great promise for innovative insights on biology mechanisms of human disease, the high-dimensionality of omics data and the complexity between various levels of omics data and disease phenotypes bring tremendous analytic challenges. To address these challenges and to facilitate ongoing multi-omics analysis, we propose a Multimodal

Functional Deep Learning (MFDL) method for high-dimensional multi-omics data analysis. MFDL model the complex relationships between genetic variants and disease phenotypes through the hierarchical structure of deep neural networks and handle high-dimensional omics data by using the functional data analysis technique. Moreover, MFDL utilizes the structure of the multimodal model to model interactions between multi-omics data. Through simulation studies and real data applications, we demonstrate the advantages of MFNN in terms of prediction accuracy as well as being robust to the high dimensionality and noise of the data.

3. An Association Test for Sequencing Data Based on Kernel Neural Networks

Tingting Hou, Department of Biostatistics, University of Florida

The recent development of artificial intelligence in genomics holds great promise to unravel the complex relationships between genetic variants and disease phenotypes and improve our understanding of the genetic etiology of complex diseases. However, due to the complexity of neural networks and their unknown limiting distributions, building significance tests on neural networks to examine complex genotype-phenotype relationships remains a great challenge. We previously developed a kernel-based neural network (KNN) method, which inherits features from both linear mixed models (LMM) and classical neural networks. Based on the KNN framework, we propose a Wald-type test to evaluate the joint association of a set of genetic variants with a disease phenotype, considering non-linear and non-additive effects. In addition, we also provide two tests to evaluate the linear genetic effect and non-linear/non-additive genetic effects (e.g., interaction effects), respectively. Through simulations, we demonstrated that our proposed method attained higher power compared to the sequence kernel association test (SKAT), especially in the presence of non-linear and interaction effects.

4. Selection of Variance Component for KNN

Heng Ge[1]*, Xiaoxi Shen[2], Qing Lu[1]
[1]Department of Biostatistics, University of Florida
[2]Texas State University

Linear mixed model (LMM) and its extensions have long been the state-of-the-art choice for genetic risk prediction analysis of complex diseases. However, LMM commonly assumes a linear genotype-phenotype relationship, which may not be satisfied for diseases involving complex genetic etiology (e.g., interactions). Moreover, it remains computationally challenging to model a large number of samples, especially with multiple kernels built on genetic data. To address these rising challenges, we propose a penalized kernel neural network (KNN) method with a minimum quadratic unbiased estimator (MINQUE). The new method uses the hierarchical structure of kernel neural networks to model complex genotype-phenotype relationships (e.g., non-linear and non-additive relationships), and a closed-form solution is derived to improve computational efficiency. We further extend Closed-form MINQUE to a numerical solution,

which can accommodate a variety of penalty terms. Through simulations and analysis PET-imaging outcomes, we have shown that penalized KNN with MINQUE estimator outperformed current state-of-the-art methods in terms of accuracy and statistical efficiency.

Contributed Session B

5. Death or Survival: Which You Measure May Affect Conclusions

Jake Shannin[1*], Babette A. Brumback[2]
[1]Department of Statistics, University of Florida
[2]Department of Biostatistics, University of Florida

Background and Aims: Considering the opposite outcome, for example, survival instead of death, may affect conclusions about which subpopulation benefits more from a treatment or suffers more from an exposure. Methods: For a case study on bankruptcy following melanoma, we compute and interpret the relative risk, odds ratio, and risk difference for different age groups. Since there is no established effect measure or outcome for either study, we redo this analysis for survival and solvency. Results: In a case study on bankruptcy following melanoma treatment, the relative risk of bankruptcy suggested that 80-89-year-old patients suffer more financially from melanoma-related expenses than 20-34-year-old patients. The relative risk of solvency and the risk difference suggested the opposite conclusion. Conclusion: To increase transparency around this paradox, researchers reporting one outcome should note if considering the opposite outcome would yield different conclusions. Researchers should also report or estimate underlying risks alongside effect measures when possible.

6. Effective Comparison of Two Potentially Crossing Hazard Rate Curves

Xiaoxi Zhang[1*], Peihua Qiu[1], Somnath Datta[1]
[1]Department of Biostatistics, University of Florida

In survival data analysis, comparison of two hazard rate curves is critically important for evaluating a treatment effect. In many applications, the two hazard curves could potentially cross each other, violating the proportional hazards assumption in Cox's model. In such cases, the traditional test like the log-rank test that was developed based on that assumption would be ineffective. There have been discussions in the literature on comparison of two potentially crossing hazard curves, based on either parametric modeling or nonparametric testing approaches. However, the assumed models of the parametric methods are difficult to justify in practice. On the other hand, the nonparametric tests are usually based on maximization with respect to an unknown crossing point, leading to complex null distributions for the corresponding test statistics. We suggest a novel method for comparing two hazard curves based on a nonparametric testing procedure. Its test statistic avoids the maximization mentioned above and consequently has the desirable asymptotic normality property under

some regularity conditions. We show that the new method is effective for comparing two crossing hazard curves.

7. A Novel Causal Mediation Analysis Approach for Zero-Inflated Mediators

Meilin Jiang, Department of Biostatistics, University of Florida

Mediation analyses play important roles in making causal inference in biomedical research to examine causal pathways that may be mediated by one or more inter- mediate variables (i.e., mediators). Although mediation frameworks have been well established such as counterfactual-outcomes (i.e., potential-outcomes) models and traditional linear mediation models, little effort has been devoted to dealing with mediators with zero-inflated structures due to challenges associated with excessive zeros. We develop a novel mediation modeling approach to address zero-inflated mediators containing true zeros and false zeros. The new approach can decompose the total mediation effect into two components induced by zero-inflated structures: the first component is attributable to the change in the mediator on its numerical scale which is a sum of two causal pathways and the second component is attributable only to its binary change from zero to a non-zero status. An extensive simulation study is conducted to assess the performance and it shows that the proposed approach outperforms existing standard causal mediation analysis approaches. We also showcase the application of the proposed approach to a real study in comparison with a standard causal mediation analysis approach.

8. An iterative method for detecting outlying studies in meta-analysis

Zhuo Meng[1*], Jingshen Wang[2], Lifeng Lin[3], Chong Wu[4]
[1]Department of Statistics, Florida State University
[2]University of California, Berkeley
[3]University of Arizona
[4]University of Texas MD Anderson Cancer Center

Meta-analysis is a widely used tool for synthesizing results from multiple studies. The collected studies are deemed heterogeneous when they do not share a common underlying effect size; thus, the factors attributable to the heterogeneity need to be carefully considered. A critical problem in meta-analyses and systematic reviews is that outlying studies are frequently included, which can lead to invalid conclusions and affect the robustness of decision-making. Outliers may be caused by several factors such as study selection criteria, low study quality, small-study effects, etc. Although outlier detection is well-studied in the statistical community, limited attention has been paid to meta-analysis. The conventional outlier detection method in meta-analysis is based on a leave-one-study-out procedure. However, when calculating a potentially outlying study, deviate, other outliers could substantially impact its result. This article proposes an iterative method to detect potential outliers, which reduces such an impact

that could confound the detection. Furthermore, we adopt bagging to provide valid inference after removing outliers. Based on simulation studies, the proposed iterative method yields smaller bias and heterogeneity after performing a sensitivity analysis of removing the identified outliers. It also provides higher accuracy on outlier detection. Two case studies are used to illustrate the proposed method in real-world performance.

Session C

9. JUMP: replicability analysis of high-throughput experiments with applications to spatial transcriptomic studies

Pengfei Lyu[1]*, Yan Li[2], Xiaoquan Wen[3], Hongyuan Cao[1]
[1]Department of Statistics, Florida State University
[2]Jilin University
[3]University of Michigan

Replicability is the cornerstone of scientific research. The current statistical method for high-dimensional replicability analysis either cannot control the false discovery rate (FDR) or is too conservative. We propose a statistical method, JUMP, for the high-dimensional replicability analysis of two studies. The input is a high-dimensional paired sequence of p-values from two studies and the test statistic is the maximum of p-values of the pair. JUMP uses four states of the p-value pairs to indicate whether they are null or non-null. Conditional on the hidden states, JUMP computes the cumulative distribution function of the maximum of p-values for each state to conservatively approximate the probability of rejection under the composite null of replicability. JUMP estimates unknown parameters and uses a step-up procedure to control FDR. By incorporating different states of composite null, JUMP achieves a substantial power gain over existing methods while controlling the FDR. Analyzing two pairs of spatially resolved transcriptomic datasets, JUMP makes biological discoveries that otherwise cannot be obtained by using existing methods.

10. Inferring cell-cell communications from spatially resolved transcriptomics data

Dongyuan Wu[1*], Jeremy Gaskins[2], Michael Sekulam[2], Susmita Datta[1]
[1]Department of Biostatistics, University of Florida
[2]University of Louisville

Cellular communication through biochemical signaling is fundamental to every biological activity. Investigating cell signaling diffusions across cell types can further help understand biological mechanisms. In recent years, it has become a rather important research topic with the maturity of single-cell sequencing technologies. However, cell signaling activities are spatially constrained, and the single-cell data cannot provide spatial information for each cell. This issue may cause a high false discovery rate, and using spatial transcriptomics data is necessary. On the other hand, as far as we know, most existing methods focus on providing an

arbitrary measurement to estimate intercellular communication instead of relying on a statistical model. It is undeniable that descriptive statistics are straightforward and more accessible, while a suitable statistical model can provide a more accurate and reliable estimate. In this way, we propose a generalized linear regression model to infer cellular communications based on spatially resolved transcriptomics data, especially spot-based data. Our Bayesian approach estimates the communication scores between cell types with the consideration of their corresponding distances. Due to the property of the regression model, the proposed method naturally provides the direction of the effect of cell types on a specific signaling pathway that other approaches cannot offer. We conduct simulation studies to assess the performance under different scenarios. We also employ the proposed model in a real-data application and compare it with other existing algorithms. In summary, our innovative model can fill in gaps in the inference of cell-cell communication and provide a robust and straightforward result.

11. Incorporating Subsampling into Bayesian Model for High-Dimensional Spatial Data

Sudipto Saha[1*], Jonathan Bradley[1]
[1]Department of Statistics, Florida State University

Additive spatial statistical models with weakly stationary process assumptions have become standard in spatial statistics. However, one disadvantage of such models is the computation time, which rapidly increases with the number of datapoints. The goal of this article is to apply an existing subsampling strategy to standard spatial additive models and to derive the spatial statistical properties. We call this new strategy the "spatial data subset model" approach, which can be applied to big datasets in a computationally feasible way. Our approach has the advantage that one does not require any additional restrictive model assumptions (i.e., computational gains increase as model assumptions are removed). This provides a solution to computational bottlenecks that occur when applying methods such as Kriging to “big data”. We provide several properties of this new "spatial data subset model" approach in terms of moments, sill, nugget, and range under several sampling designs. We present the results of the "spatial data subset model" approach on a simulated dataset, and on a large dataset consists of 150,000 observations of daytime land surface temperatures measured by the MODIS instrument onboard the Terra satellite.

12. Modeling of clustered censored survival data with spatial correlated random effects using SBART

Durbadal Ghosh[1*], Dr Debajyoti Sinha[1], Jonathan Bradley[1], George Rust[1]
[1]Department of Statistics, Florida State University

This article presents a nonparametric Bayesian survival analysis of clustered survival data where cluster effects are spatially correlated, and some of the spatially-correlated cluster-level

covariates are estimated using a data source different from the survival data source. Widespread parametric and semi-parametric hazards regression models for clustered survival data are inappropriate and inadequate for this situation, with complex spatial effects on both clustering and covariates. In this article, we present a general nonparametric model for such clustered censored survival data with spatial correlation under a paradigm of Bayesian ensemble learning called Soft Bayesian Additive Regression Trees or SBART. Our computationally feasible method can incorporate unknown functional forms of the main effects and interactions of various covariates and cluster-specific effects. We illustrate the practical implementation of our method with an analysis of the effects of various intervenable and non-intervenable covariates on survival times of breast cancer patients from different counties (clusters) in Florida. For our analysis, the clustered survival data with patient-level covariates come from Florida Cancer Registry (FCR), and the data for one county-level intervenable covariates come from the Behavioural Risk Factor Surveillance Survey (BRFS). We compare our method with existing analysis methods to demonstrate our advantage in assessing the impacts of intervention in some cluster/county level and patient-level covariates to eliminate racial disparity in breast-cancer survival in different Florida counties.

Session D

13. Skew-Normal Classification in High-Dimensional data

Haesong Choi[1*] and Qing Mai[1]
[1]Department of Statistics, Florida State University

A considerable number of studies have been devoted to high-dimensional classification models under the assumption of normality. However, it may be too restrictive in data analysis. Motivated by the data set that exhibits asymmetry, including environmental, financial, and biomedical ones, we propose a high-dimensional discriminant analysis model called the SKNC model (short for SKew-Normal Classification). By incorporating the skew-normal model, the SKNC model inherits convenient formal properties of the normal distribution and improves its flexibility on skewed data in classification. Theoretical results rigorously show that the SKNC model achieves variable selection, penalized estimation, and prediction consistency, especially in high-dimensional settings. We empirically demonstrate the superior performance of the SKNC model over existing methods in simulated and real datasets.

14. High-dimensional Posterior Consistency in Multi-response Regression models with Non-informative Priors for Error Covariance Matrix

Partha Sarkar[1*], Kshitij Khare[1], Malay Ghosh[1]
[1]Department of Statistics, University of Florida

The Inverse-Wishart (IW) distribution is a standard and popular choice of priors for covariance matrices and has attractive properties such as conditional conjugacy. However, the IW family of priors has crucial drawbacks, including the lack of effective choices for non-informative priors. Several classes of priors for covariance matrices that alleviate these drawbacks, while preserving computational tractability, have been proposed in the literature. These priors can be obtained through appropriate scale mixtures of IW priors. However, the high-dimensional posterior consistency of models which incorporate such priors has not been investigated. We address this issue for the multi-response regression setting (q responses, n samples) under a wide variety of IW scale mixture priors for the error covariance matrix. Posterior consistency and contraction rates for both the regression coefficient matrix and the error covariance matrix are established in the ``large q, large n'' setting under mild assumptions on the true data-generating covariance matrix and relevant hyperparameters. In particular, the number of responses q_n is allowed to grow with n, but with $q_n = o(n)$. Also, some results related to the inconsistency of posterior mean for $q_n/n \to \gamma$, where $\gamma \in (0, \infty)$ are provided.

15. Multivariate single index modeling of longitudinal data with multiple responses

Zibo Tian[1*] and Peihua Qiu[1]
[1]Department of Biostatistics, University of Florida

In medical studies, composite indices and/or scores are routinely used for predicting medical conditions of patients. These indices are usually developed from observed data of certain disease risk factors, and it has been demonstrated in the literature that single index models can provide a powerful tool for this purpose. In practice, the observed data of disease risk factors are often longitudinal in the sense that they are collected at multiple time points for individual patients, and there are often multiple aspects of a patient's medical condition that are of our concern. However, most existing single index models are developed for cases with independent data and a single response variable, which are inappropriate for the problem just described in which within-subject observations are usually correlated and there are multiple mutually correlated response variables involved. This paper aims to fill this methodological gap by developing a single index model for analyzing longitudinal data with multiple responses. Both theoretical and numerical justifications show that the proposed new method provides an effective solution to the related research problem. It is also demonstrated using a dataset from the English Longitudinal Study of Aging.

16. A general framework for online monitoring of dynamic networks

Yipeng Wang[1*], Xiulin Xie[1], and Peihua Qiu[1]
[1]Department of Biostatistics, University of Florida

Dynamic networks where nodes or edges are in constant change become increasingly common due to advances in information technology and data collection systems. A dynamic network process takes the form of a data stream in the sense that new networks keep being collected

over time. One fundamental research problem is to monitor the underlying sequential process of the observed networks to see whether it is longitudinally stable, and a signal should be given as soon as an anomaly occurs. To deal with an anomalous change without specializing change types, we propose (a) characterizing each network by correlated summary statistics from the whole network or connected components, and (b) using a nonparametric multivariate cumulative sum (CUSUM) chart to monitor the serially correlated process of the summary statistics. If nodes keep changing in the process, the number of nodes, average degree, the number of connected components, and average diameter are monitored simultaneously. As a special case, if nodes are fixed, the aforementioned summary statistics except the number of nodes are used to be monitored. Traditionally, a monitoring strategy is based on a fixed node set and assumptions for the dynamic network model, which may not be valid in practice. For example, fitting a stochastic block model that requires network communities remain unchanged is questionable when dynamic networks are constantly changing. As a result, monitoring the model-based metrics is dubious. In addition, existing methods based on conventional control charts suffer from unrealistic assumptions of normality or time-independent. The new general framework is robust to use if these assumptions are violated. A procedure based on the Poisson hurdle model is proposed to generate sparse dynamic networks rather than simple random graph models. Numerical studies show that our monitoring scheme is generally more reliable and effective than the representative existing methods.