2024 Annual Meeting of the American Statistical Association Florida chapter

March 29-30, 2024

Hosted by Florida State University

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

FLORIDA STATE UNIVERSITY
VIRES ARTES MORES
1851

Welcome to Florida State University!

The Florida State University Department of Statistics is delighted to host this year's meeting of the Florida Chapter of the American Statistical Association! We hope you enjoy your time in Tallahassee, Florida! Meals are not provided here at the meeting, but Tallahassee has many local restaurants that serve delicious food. Here are some restaurants that are popular with us:

Bella Bella

If Italian food is what you're craving, you'll definitely want to check out Bella Bella. Located at 123 East 5th Avenue, they have a fun atmosphere that adds an extra dimension to your meal. Their specialty is bubble bread, so be sure to try some!

Masa

Owned by local pillar of the community, Lucy Ho, Masa is one of her many restaurants that focuses on Asian-fusion cuisine. Located at 1650 North Monroe Street, it's worth the short drive! Masa offers a variety of sushi, noodle, and rice dishes that will make anyone's mouth water.

Cafe de martin

Home to Peruvian Flavors in Tallahassee! Led by Chef Martín, originally from Lima, Peru, he's renowned in Leon County's Hispanic community for his outstanding Peruvian cuisine.

Harry's Seafood Bar and Grille

Harry's of Tallahassee is the place for Cajun, Creole, and Southern-style cooking. They are located at 301 South Bronough Street, and definitely worth the drive. Their wide menu selection means that there's truly something for everyone!

Savour

A hub of hospitality, is the chic and eclectic dining experience in downtown Tallahassee. With a vision for seasonally inspired, regionally sourced and creatively prepared cuisine, Chef Brian Knepper brings more to the table even for the most discriminating gastronomes.

Overview of Events

Friday March 29th

9:00 AM – 9:45 AM	Registration & Breakfast
9:45 AM – 10:00 AM	Opening & Welcome
10:00 AM - 11:30 AM	Invited session A – AI and Machine learning
11:30AM – 1:00 PM	Lunch break
1:00 PM – 2:15 PM	Invited Panel Discussion – Career development
2:15 PM – 2:30 PM	Coffee break
2:30 PM – 4:20 PM	Parallel sessions
4:30 PM – 5:30 PM	Student posters
6:00 PM – 6:50 PM	Keynote from Dr. Ji-Hyun Lee
6:50 PM – End of the day	Dinner Banquet

Saturday March 30th

8:00 AM – 9:00 AM	Business meeting of FL chapter
9:00 AM – 10:00 AM	Keynote from Dr. Hulin Wu
10:00 AM - 10:30 AM	Coffee break
10:30AM – 12:00 PM	Invited session B – Biostatistics and Bioinformatics
12:00 PM – 12:15 PM	Closing remarks, Student awards ceremony

Friday, March 29th

9:00am – 9:45am Registration

9:45am – 10:00am Opening & Welcome: **Eric Chicken** (Organizer & Chair: Xin (Henry) Zhang), **FLH 0275**

10:00am – 11:30am **Invited session A – AI and Machine learning** (Organizer & Chair: Chao Huang)**, FLH 0275**

10:00am – 10:30am **Hsin-Hsiung Huang**, University of Central Florida

From algorithms for anomaly detection to spatial and temporal modeling and Bayesian ultra-high dimensional variable selection

Inspired by our investigation on spatiotemporal data analysis for the NSF ATD challenges, we've investigated Bayesian clustering, variable selection for mixed-type multivariate responses and Gaussian process priors for spatiotemporal data. The proposed Bayesian approaches effectively and efficiently fit high-dimensional data with spatial and temporal features. We further propose a two-stage Gibbs sampler which leads a consistent estimator with a much faster posterior contraction rate than a one-step Gibbs sampler. For Bayesian ultrahigh dimensional variable selection, we have developed Bayesian sparse multivariate regression for mixed responses (BS-MRMR) with shrinkage priors model for mixed-type response generalized linear models. We consider a latent multivariate linear regression model associated with the observable mixed-type response vector through its link function. Under our proposed BS-MRMR model, multiple responses belonging to the exponential family are simultaneously modeled and mixed-type responses are allowed. We show that the MBSP-GLM model achieves posterior consistency and quantifies the posterior contraction rate. Additionally, we incorporate Gaussian processes into zero-inflated negative binomial regression. To conquer the computation bottleneck that GPs may suffer when the sample size is large, we adopt the nearest-neighbor GP approach that approximates the covariance matrix using local experts. We provide simulation studies and real-world gene data examples.

Muxuan Liang, University of Florida

A General Framework for Incorporating Identification Uncertainty in Individualized Treatment Rules

Estimating individualized treatment rules (ITRs) from observational data or clinical trials with non-adherence is challenging due to possible unmeasured confounding bias. Partial identification approaches using an instrumental variable (IV) provide characterizations on possible values of the conditional average treatment effects (CATEs). In this work, we develop a new class of `optimal' ITRs to guide treatment decisions when the CATEs are only partially identified. We define a novel value function allowing a reject option in treatment decisions under partial identification, and use that value function to define a class of IV-optimal ITRs with a reject option. The reject option informs who are susceptible to identification uncertainty and allows the use of alternative ITRs derived from other studies or outcomes for these

patients. In addition, our framework allows users to control the size of subgroups receiving the reject option, taking into account the risks associated with unreliable or delayed treatment assignments. To estimate the IV-optimal ITRs with a reject option, we develop a weighted classification framework with a modified hinge loss function, where the weights are non-smooth transformations of nuisance parameters. We further propose an empirical augmented risk minimization approach that achieves a fast convergence rate even if the nuisance parameters are estimated using nonparametric or machine learning methods. Simulations and real data analysis are conducted to demonstrate the superiority of the developed framework and estimation procedure.

Jiawei Zhang, University of Kentucky

Is a Classification Procedure Good Enough?—A Goodness-of-Fit Assessment Tool for Classification Learning

In recent years, many nontraditional classification methods, such as random forest, boosting, and neural network, have been widely used in applications. Their performance is typically measured in terms of classification accuracy. While the classification error rate and the like are important, they do not address a fundamental question: Is the classification method underfitted? For a general classification procedure, the lack of a parametric assumption makes it challenging to construct proper tests. To overcome this difficulty, we propose a methodology called BAGofT that splits the data into a training set and a validation set. First, the classification procedure to assess is applied to the training set, which is also used to adaptively find a data grouping that reveals the most severe regions of underfitting. Then, based on this grouping, we calculate a test statistic by comparing the estimated success probabilities and the actual observed responses from the validation set. The data splitting guarantees that the size of the test is controlled under the null hypothesis, and the power of the test goes to one as the sample size increases under the alternative hypothesis. For testing parametric classification models, the BAGofT has a broader scope than the existing methods since it is not restricted to specific parametric models (e.g., logistic regression). Extensive simulation studies show the utility of the BAGofT when assessing general classification procedures and its strengths over some existing methods when testing parametric classification models.

11:30am – 1:00pm Lunch break

1:00pm – 2:15pm **Panel Discussion-Career development** (Organizer & Moderator: Joshua Loyal), **FLH 0275**

Panelists:

Dereck Tucker, Distinguished Member of the Technical Staff, Statistical Sciences, Sandia National Laboratories

Michiko Wolcott, Partner and Principal Consultant, Msight Analytics

Nathan Crock, Director at NewSci Labs and Affiliate Faculty in the Department of Scientific Computing and the Interdisciplinary Data Science Program at Florida State University

BeeJay Girimurugan, Associate Professor in the Department of Mathematics and Interim Associate Dean of The Honors College at Florida Golf Coast University

2:00pm – 2:30pm Break

2:30pm – 3:20pm **Parallel Session A1** (Chair: Joshua Loyal), **FLH 0275**

2:30pm – 2:40pm **Peak p-values for Gaussian Random Fields on A Lattice**

Tuo Lin, University of Florida

In this work we develop a Monte Carlo method to compute the height distribution of local maxima of a stationary Gaussian or Gaussian derived random field that is observed on a regular lattice. We show that our method can be used to provide valid peak based inference in order to detect and localize signals in random fields, under minimal assumptions. This offers a solution for small values of FWHM when the existing formula derived for continuous domain is not accurate. We also extend the methods in Worsley (2005) and Taylor et al. (2007) to compute the height distribution and compare them with our approach. Lastly, we apply our method to a task fMRI dataset to show how it can be used in practice.

2:40pm – 2:50pm **Parametric Bootstrap and Fiducial Inference for Two-parameter Maxwell Distributions**

Faysal Chowdhury, Florida Gulf Coast University

The two-parameter Maxwell distribution is getting popular as one of the lifetime distributions due to its smoothly increasing failure rate. Our study focuses on constructing confidence intervals (CIs) for the difference between means and ratio of means of two independent Maxwell distributions. We propose CIs based on the fiducial approach, approximate fiducial approach (also known as modified normal-based approximation), and parametric bootstrap (PB) method. We compare these methods based on their coverage probability and precision. We extend these methods to find CIs for a difference between percentiles and a ratio of two independent Maxwell distributions. Specifically, we develop and evaluate CIs for the ratio of the 5th percentiles and the ratio of the medians for coverage probability and precision accuracy. We illustrate these methods using two examples with real-life data. Although the PB confidence intervals are more efficient than the fiducial CIs in some situations, the approximate fiducial CIs are very simple to compute and are comparable with the PB CIs in most cases.

2:50pm – 3:00pm **Replicability analysis of high dimensional data accounting for dependence**

Pengfei Lyu, Florida State University

Replicability is the cornerstone of scientific research. We study the replicability of data from high-throughput experiments, where tens of thousands of features are examined simultaneously. Existing replicability analysis methods either ignore the dependence among features or impose strong modeling assumptions, producing overly conservative or overly liberal results. Based on two sequences of p-values, we use a four-state hidden Markov model to capture the structure of local dependence in two heterogeneous studies. Our method effectively borrows information from different features and studies while accounting for dependence among features and heterogeneity across studies. We show that the proposed method has better power than competing methods while controlling the false discovery rate, both empirically and theoretically. Analyzing datasets from the genome-wide association studies reveals new biological insights that otherwise cannot be obtained by using existing methods.

3:00pm – 3:10pm

Analysis of spatially clustered survival data with unobserved covariates using SBART

Durbadal Ghosh, Florida State University

Popular parametric and semi-parametric regression methods for clustered survival data are inappropriate and inadequate when the appropriate functional forms of the covariates and their interactions are unknown, and random cluster effects as well as some unknown cluster-level covariates are spatially correlated. We present a general nonparametric method for such data under the Bayesian ensemble learning paradigm called soft Bayesian additive regression trees or SBART in short. Our additional methodological and computational challenges include a large number of clusters, variable cluster sizes, and data information for proper statistical augmentation of the unobserved covariate sourced from a data registry different from the survival study.

We illustrate the practical implementation of our method and its advantages over existing methods by assessing the impacts of intervention in some cluster/county level and patient-level covariates to eliminate existing racial disparity in breast cancer survival in different Florida counties (clusters), where the clustered survival data with patient-level covariates come from Florida Cancer Registry (FCR), and the data information for an unobserved county-level covariate comes from the Behavioral Risk Factor Surveillance Survey (BRFSS). We also compare our method with existing analysis methods to demonstrate its advantages through simulation studies.

3:10pm – 3:20pm

Multiblock Partial Least Squares and Rank Aggregation: Applications to Detection of Bacteriophages Associated with Antimicrobial Resistance in the Presence of Potential Confounding Factors

Shoumi Sarkar, University of Florida

Urban environments, characterized by bustling mass transit systems and high population density, host a complex web of microorganisms that impact microbial interactions. These urban microbiomes, influenced by diverse demographics and constant human movement, are vital for understanding microbial dynamics. We explore urban metagenomics, utilizing an extensive dataset from the Metagenomics & Metadesign of Subways & Urban Biomes (MetaSUB) consortium, and investigate antimicrobial resistance (AMR) patterns. In this pioneering research, we delve into the role of bacteriophages, or "phages" - viruses that prey on bacteria and can facilitate the exchange of antibiotic resistance genes (ARGs) through mechanisms like horizontal gene transfer (HGT). Despite their potential significance, existing literature lacks a consensus on their significance in ARG dissemination. We argue that they are an important consideration. We uncover that environmental variables, such as those on climate, demographics, and landscape, can obscure phage-resistome relationships. We adjust for these potential confounders and clarify these relationships across specific and overall antibiotic classes with precision, identifying several key phages. Leveraging machine learning tools and validating findings through clinical literature, we uncover novel associations, adding valuable insights to our comprehension of AMR development.

2:30pm – 3:20pm

Parallel Session B1 (Chair: Zhaotong Lin), OSB 0108

2:30pm – 2:40pm

A New Framework for Bayesian Function Registration

Wei Wu, Florida State University

Function registration, also referred to as alignment, has been one of the fundamental problems in the field of functional data analysis. Classical registration methods such as Fisher-Rao alignment focus on estimating optimal time warping function between functions. In recent studies, the model on time warping has attracted more attention, and it can be used as a prior term to combine with the classical method (as a likelihood term) in a Bayesian framework. The Bayesian frameworks have been shown improvement over the

classical approaches. However, its prior model on time warping is often based a nonlinear approximation, which may introduce inaccuracy, instability, and inefficiency. To overcome these problems, we propose a new Bayesian approach by adopting a prior which provides a linear representation and various stochastic processes (Gaussian or non-Gaussian) can be effectively utilized on time warping. No approximation is needed in the time warping computation, and posterior distribution can be obtained via a conventional Markov chain Monte Carlo approach. We thoroughly investigate the impact of prior on the performance of functional registration with multiple simulation examples, which demonstrate the superiority of the Bayesian framework over the classical methods. We finally utilize the new method in a real dataset and obtain desirable alignment result.

2:40pm – 2:50pm

Comparing Two Hazard Curves When There Is a Treatment Time-lag Effect

Xiaoxi Zhang, University of Florida

In cancer and other medical studies, time-to-event (e.g., death) data are common. One major task to analyze time-to-event (or survival) data is usually to compare two medical interventions (e.g., a treatment and a control) regarding their effect on patients' hazard to have the event in concern. In such cases, we need to compare two hazard curves of the two related patient groups. In practice, a medical treatment often has a time-lag effect, i.e., the treatment effect can only be observed after a time period since the treatment is applied. In such cases, the two hazard curves would be similar in an initial time period, and the traditional testing procedures, such as the log-rank test, would be ineffective in detecting the treatment effect because the similarity between the two hazard curves in the initial time period would attenuate the difference between the two hazard curves that is reflected in the related testing statistics. In this paper, we suggest a new method for comparing two hazard curves when there is a potential treatment time-lag effect based on a weighted log-rank test with a flexible weighting scheme. The new method is shown to be more effective than some representative existing methods in various cases when a treatment time-lag effect is present.

2:50pm – 3:00pm

Bayesian density estimation on the product of simplexes and the hypercube using multivariate Bernstein polynomial

Rufeng Liu, Florida State University

We propose a Bayesian nonparametric model for density estimation on the product of simplex spaces and the hypercube. The model is particularly useful for cases where the available data consist of multiple compositional features alongside variables that take on values within bounded intervals. A compositional feature is a vector of non-negative components whose sum of values remains constant, such as the time an individual spends on different activities during the day or the fraction of different types of food consumed as part of a person's diet. Our approach relies on a generalization of random multivariate Bernstein polynomials and corresponds to a Dirichlet process mixture of products of Dirichlet and beta densities. Theoretical properties such as prior support and posterior consistency are studied. We evaluate the model's performance through a simulation study and a real-world application using data from the 2005–2006 cycle of the U.S. National Health and Nutrition Examination Survey (NHANES). Furthermore, the conditional densities derived under this modeling strategy can be used for regression analyses where both the response and predictors take values on the simplex space and/or hypercube.

3:00pm – 3:10pm

Geo-DMAE: Integrative Clustering and Reconstruction of Multiple Subcortical Brain Structures using Geometric Deep Multi-Autoencoders

Yuanyao Tan, Florida State University

Multiple subcortical brain structure alternations (such as shape alterations in the hippocampus and amygdala) are typically interrelated and connected with the natural process of brain aging. Detecting and understanding the geometric differences in these diverse subcortical brain structures is essential for monitoring brain aging. However, existing learning methods face several hurdles in modeling these variations, including (i) subgroup-

level imaging heterogeneity caused by distinct brain aging mechanisms, (ii) integration of multiple 2D/3D subcortical shapes, and (iii) optimal Riemannian local embedding representations of shapes across different subgroups.
To address these challenges, we propose a Geometric Deep Multi-Autoencoders (Geo-DMAE) framework to simultaneously detect the latent subgroup patterns within the subjects and reconstruct multiple subcortical brain structures at both subject-level and subgroup-level. Specifically, the subcortical shapes are first derived by removing the translation, scaling, and rotation variabilities and projected to hyper-spherical spaces. In each geometric deep autoencoder, the prealigned shapes are fed through a rotation \& embedding layer and a multilayer perceptron (MLP) based autoencoder, where the subgroup patterns are learned via fitting incremental Gaussian mixture models to the latent features integrated from multiple autoencoders. Finally, both simulation studies and real data analysis based on multiple brain subcortical structures from the Alzheimer's disease study are conducted to evaluate the finite sample performance of Geo-DMAE.

3:10pm – 3:20pm **Interval-specific censoring set adjusted Kaplan–Meier estimator**

Yaoshi Wu, Cytokinetics

Interval-specific censoring set adjusted Kaplan–Meier estimator (WKE) is a non-parametric approach to reduce the overestimation of the Kaplan-Meier estimator (KME) when the event and censoring times are independent. The article is published in J. Appl. Stat1. We adjusted the KME based on a collection of intervals where censored data are observed between two adjacent event times. WKE is superior to KME as it substantially reduces the overestimation on survival rate and median survival time in the presence of censoring data. When there are no censored observations, or the sample size goes to infinity, WKE reduces to KME. We proved theoretically that WKE reduces overestimation compared to KME and provided a mathematical formula to estimate the variance of the proposed estimator based on Greenwood's approach. We performed four simulation studies to compare WKE with KME when the failure rate is constant, decreasing, increasing, and based on the flexible hazard method. The bias reduction in median survival time and survival rate using WKE is considerably large, especially when the censoring rate is high. The standard deviations are comparable between the two estimators. We implemented WKE and KME for the Nonalcoholic Fatty Liver Disease patients from a well conducted population study. The results based on the actual data also show WKE substantially reduces the overestimation in the presence of high observed censoring rate.
1 Yaoshi Wu & John Kolassa (25 Dec 2023): Interval-specific censoring set adjusted Kaplan–Meier estimator, Journal of Applied Statistics, DOI: 10.1080/02664763.2023.2298795

3:20pm – 3:30pm Break

3:30pm – 4:20pm **Parallel Session A2** (Chair: Chao Huang), **FLH 0275**

3:30pm – 3:40pm **Kernel meets sieve: transformed hazards models with sparse longitudinal covariates**

Hongyuan Cao, Florida State Univerisity

We study the transformed hazards model with time-dependent covariates observed intermittently for the censored outcome. Existing work assumes the availability of the whole trajectory of the time-dependent covariates, which is unrealistic. We propose combining kernel-weighted log-likelihood and sieve maximum log-likelihood estimation to conduct statistical inference. The method is robust and easy to implement. We establish the asymptotic properties of the proposed estimator and contribute to a rigorous theoretical framework for general kernel-weighted sieve M-estimators. Numerical studies corroborate our theoretical results and show that the proposed method performs favorably over competing methods. The analysis of a data set from a COVID-19 study in Wuhan identifies clinical predictors that otherwise cannot be obtained using existing methods.

3:40pm – 3:50pm

Input-Response Space Filling Designs Incorporating Response Uncertainty

Xiankui Yang, University of South Florida

Traditionally space-filling designs have focused on the characteristics of the design in the input space ensuring uniform spread throughout the region. Input-Response Space Filling designs considered scenarios when having good spread throughout the range or region of the responses is also of interest. This paper acknowledges that there is typically uncertainty associated with the values of the response(s) and hence proposes a method, Input-Response Space Filling Designs with Uncertainty (IRSFwU), to incorporate this into the design construction. The Pareto front of designs offers alternatives that balance input and response space filling, while prioritizing input combinations with lower associated response uncertainty. These lower uncertainty choices improve the chances of observing the desired response value. We describe the new approach with an uncertainty-adjusted distance to measure the response space filling, the Pareto aggregate point exchange algorithm to populate the set of promising designs, and illustrates the method with three examples of different input and response relationships and dimensions.

3:50pm – 4:00pm

Risk Monitoring of Multiple Diseases by Flexible Modeling of Longitudinal Data

Zibo Tian, University of Florida

Disease early detection and prevention are important topics in medical and health research. One application of sequential monitoring of dynamic processes, the dynamic disease screening system (DySS), is a powerful tool for giving early signals of individuals' irregular longitudinal pattern in one or multiple predictors of a target disease. In practice, multiple diseases (e.g., different types of cardiovascular disease) may need to be monitored and prevented simultaneously. In this paper, a new DySS is proposed to jointly monitor the risk of multiple diseases. Given a training dataset, we fit a multivariate single-index logistic regression model with random effects and quantify the multivariate risk of developing the target diseases for each subject at the corresponding measurement time points. Then, the longitudinal pattern of the underlying risk of those well-functioning subjects is estimated. For an incoming subject, a signal will be triggered by a large cumulative difference between the longitudinal risk pattern of the current subject and the regular longitudinal pattern. Numerical studies show that the proposed method works well in different scenarios and is more powerful than using the conventional multivariate DySS to monitor the pool of predictors for the multiple diseases of interest.

4:00pm – 4:10pm

De-correlated Nearest Shrunken Centroids for Tensor Data

Munwon Yang, Florida State University

The nearest shrunken centroids (NSC) method is an efficient and accurate classifier. However, it is challenged by the recently prevalent tensor data, as it does not model the correlation structure among the tensor predictors. We tackle this challenge by proposing a new distance-based classifier, TDNSC. TDNSC leverages the popular separable covariance structure on tensor data to decorrelate data and allow easy application of NSC afterwards. The theoretical properties and empirical results suggest that TDNSC is a promising method for tensor classification.

3:30pm – 4:20pm

Parallel Session B2 (Chair: Jonathan Stewart), OSB 0108

3:30pm – 3:40pm

Differentially Private Methods for Managing Model Uncertainty in Linear Regression Models

A. Felipe Barrientos, Florida State University

Many data producers are concerned about protecting individuals' private information while still allowing modelers to draw inferences from confidential data sets. The framework of differential privacy enables statistical analyses while controlling and quantifying the potential leakage of private information. In this talk, we present differentially private methods for

hypothesis testing, model averaging, and model selection for normal linear models. We consider both Bayesian and non-Bayesian methods for the tasks. The procedures are asymptotically consistent and straightforward to implement with existing software. We focus on practical issues such as quantifying the uncertainty introduced by the privacy-ensuring mechanisms. We evaluate the empirical performance of the approaches using simulated and real data.

3:40pm – 3:50pm

Factors associated with failure to receiving standard recommended screening for colorectal cancer among respondents aged 45-75 in ohio, united state

Opeyemi Oyekola Ogungbola, Florida A&M University

This study explores the factors that contribute to respondents in Ohio, United States, between the ages of 45 and 75 who do not undergo the routinely advised screening for colorectal cancer (CRC). The study examines demographic and health-related factors using data from 2015 and 2022 and descriptive statistics, correlation matrices, chi-square tests, and logistic regression analysis. Key findings show that temporal shifts, race, gender, age, BMI, education, married status, health insurance, and urban residence all affect the recommendations for CRC tests. There are differences: those who are married, live in a city, have health insurance, are female, older, of normal weight, are educated, do not smoke, and are not Hispanic are more likely to receive recommendations. Notably, the data shows that recommendations relating to education and age have changed over time. The intricate interactions between predictors in the research are revealed by the correlation matrix. Racial differences are suggested by chi-square tests, with a substantial connection in 2022. Model validation prefers the 2015 model, indicating relationships that have changed over time, while logistic regression analysis highlights factors influencing CRC test recommendations. Recommendations for focused public health initiatives targeting race, gender, age, marital, urban, and educational inequities are provided in the study's conclusion. Future study ideas to support equal access to preventive healthcare services include health literacy initiatives, qualitative investigations, longitudinal analyses, and regional analyses.
Keywords: colorectal cancer screening, demographic factors, health literacy, preventive healthcare, and public health interventions.

3:50pm – 4:00pm

Distribution-on-scalar Single-index Quantile Regression Model for Handling Tumor Heterogeneity

Shengxian Ding, Florida State University

This paper develops a distribution-on-scalar single-index quantile regression modeling framework to investigate the relationship between cancer imaging responses and scalar covariates of interest while tackling tumor heterogeneity. Conventional association analysis methods typically assume that the imaging responses are well-aligned after some preprocessing steps. However, this assumption is often violated in practice due to imaging heterogeneity. Although some distribution-based approaches are developed to deal with this heterogeneity, major challenges have been posted due to the nonlinear subspace formed by the distributional responses, the unknown nonlinear association structure, and the lack of statistical inference. Our method can successfully address all the challenges. We establish both estimation and inference procedures for the unknown functions in our model. The asymptotic properties of both estimation and inference procedures are systematically investigated.
The finite-sample performance of our proposed method is assessed by using both Monte Carlo simulations and a real data example on brain cancer images from TCIA-GBM collection.

4:00pm – 4:10pm

Deep5hmC: Predicting genome-wide 5-Hydroxymethylcytosine landscape via multimodal deep learning model

Xin Ma, University of Florida

5-hydroxymethylcytosine (5hmC), an important epigenetic mark, which plays an important role in regulating gene expression in a tissue/cell-type specific manner, is critical for

understanding the functional dynamics of the genome. By leveraging tissue-specific 5hmC sequencing data, we develop Deep5hmC, which is a multimodal deep learning framework that integrates both the DNA sequence and the histone modification, to predict genome-wide 5hmC modification. Benefiting from the multimodal design, Deep5hmC shows remarkable improvement in predicting both qualitative and quantitative 5hmC modification compared to unimodal versions of Deep5hmC and state-of-the-art machine learning methods across 4 times points in forebrain organoid development as well 17 human tissues. Importantly, Deep5hmC demonstrates its practical usage by accurately predicting gene expression and differentially hydroxymethylated regions in a case-control Alzheimer's disease study.

4:30pm – 5:30pm **Student posters, Chemistry building lobby**

Extrinsic Principal Component Analysis, Ka Chun Wong, Florida State University

We aims to develop a methodology for extrinsic principal components. Instead of evaluating directly on object manifolds with intrinsic metric, we proposed a method that working on an embedding manifolds with extrinsic metric. This method help us analyzing object shape space with a different perspective. We define the extrinsic principal sub-manifolds of a random object on a manifold embedded in an Euclidian space, and their sample counterparts. These submanifolds are necessary for dimension data reduction of high dimensional objects such as shapes of contour data, projective shapes, 3D Kendall shapes. For applications, we retain reasonable small number of such extrinsic principal submanifolds for different data samples, extracted from imaging data.

Persitent Landscapes for Object Data, Tingan Chen, Florida State University

Persistent homology is a modern Mathematical Statistics concept, capturing the topological features of a dataset of objects representable as points on a metric space, from their coarse to fine representations. The birth and death diagrams of features such as "connected components", "holes" and higher dimensional "voids" with respect to a filtration, show the span of the scales on which these features appear, persist (be of relevance), and vanish, hence the name persistent homology. Persistent landscape, as a recent advancement, was proposed by Bubenik(2014) as a linearization of the original persistent diagram method, on which statistical analysis on vector spaces could be applied. On a toy dataset, we show that the persistent landscape can differentiate between data points sampled from two different random objects. In particular, from this perspective the current analysis of networks can be improved in certain aspects using TDA.

scaDA: A Novel Statistical Method for Differential Analysis of Single-Cell Chromatin Accessibility Sequencing Data, Fengdi Zhao, University of Florida

Single-cell ATAC-seq sequencing data (scATAC-seq) has been widely used to investigate chromatin accessibility on the single-cell level. One important application of scATAC-seq data analysis is differential chromatin accessibility (DA) analysis. However, the data characteristics of scATAC-seq such as excessive zeros and large variability of chromatin accessibility across cells impose a unique challenge for DA analysis. Existing statistical methods focus on detecting the mean difference of the chromatin accessible regions while overlooking the distribution difference. Motivated by real data exploration that distribution difference exists among cell types, we introduce a novel composite statistical test named "scaDA", which is based on zero-inflated negative binomial model (ZINB), for performing differential distribution analysis of chromatin accessibility by jointly testing the abundance, prevalence and dispersion simultaneously. Benefiting from both dispersion shrinkage and iterative refinement of mean and prevalence parameter estimates, scaDA demonstrates its superiority to both ZINB-based likelihood ratio tests and published methods by achieving the highest power and best FDR control in a comprehensive simulation study. In addition to demonstrating the highest power in three real sc-multiome data analyses, scaDA successfully identifies differentially accessible regions in microglia from sc-multiome data for

an Alzheimer’s disease (AD) study that are most enriched in GO terms related to neurogenesis and the clinical phenotype of AD, and AD-associated GWAS SNPs.

Novel Cure Rate Models Based on Proliferation of Latent Risks, Pitshou Nzazi Duki, Florida State University

We consider right-censored survival data methods for populations with a cured fraction. Unlike the existing models, we propose two classes of competing latent risk models that allow the proliferation of latent risks over time. The first class is derived from an ordinary differential equation, and the second family is based on a homogeneous Poisson process. In our work, we first describe our parametric models and demonstrate that the models have critical practical advantages, including a meaningful biological interpretation and relative ease of Bayesian implementation via drawing posterior samples using Markov Chain Monte Carlo (MCMC) methods. Second, we present some tools for the Bayesian estimati, Florida State University

on of the proposed models. Then, we propose statistical metrics for model comparison based on the Cox–Snell and the Schoenfeld residuals. Simulation studies evaluate the model performance by discussing the obtained results. Furthermore, we illustrate the practical advantages and applications of the proposed models within the frequentist and Bayesian approaches to perform inferences. We associate a residual-based analysis of data for breast cancer in the Surveillance Epidemiology and End Results (SEER) database. The results show that our models fit the data better according to the analysis.

Variable Screening and Spatial Smoothing in Fr\'echet Regression with Application to Diffusion Tensor Imaging, Lei Yan, Florida State University

Modern applications in medical imaging often include high-dimensional predictors and spatially dependent responses in the non-Euclidean space. For example, in imaging-genetics studies, our objective is to study the relationship between single-nucleotide polymorphisms (SNPs), a high-dimensional predictor vector, and diffusion tensor imaging (DTI) responses, which are thousands to millions of voxel-wise 3×3 symmetric positive definite (SPD) matrices. In this paper, We develop a fast and pragmatic method of regressing spatially associated random responses on a high-dimensional predictor set. Specifically, we focus on two related problems: fast variable screening of high-dimensional predictors and smoothing techniques for non-Euclidean spatially associated responses. Under a Fr\'echet regression framework (which handles regression of SPD matrix-variate responses on covariates in Euclidean space), we propose a two-stage approach, where a screening method (using distance covariances in metric spaces) is employed to mitigate high-dimensionality (Stage 1), followed by deriving a closed-form solution that powers elegant smoothing of the spatially associated SPD responses (Stage 2). We investigate the finite-sample properties of our method using synthetic data generated under various settings, and present illustration via analysis of an imaging-genetics (DTI responses with genetic and demographic predictors) dataset, derived from the Alzheimer's Disease Neuroimaging Initiative 2. Code for implementing our proposed method is available in the GitHub link: https://github.com/leiyan-ly/Frechet-regression."

Tensor Response Regression with Low Tubal Rank, Jiping Wang, Florida State University

For contemporary scientific data, complicated structured tensor data with high dimensions are everywhere. Motivated by modeling the relationship between the multivariate covariates with complicated tensor response, we proposed a tensor response model with low tubal rank assumption. The low tubal rank constraint can efficiently reduce the number of free parameters and its meaning is also easy to interpret. One special case of our model is equivalent to multivariate reduced rank regression model. We also put forward a proven convergent ADMM algorithm that can obtain the estimation efficiently. Simulations show that our method outperforms existing tensor response model significantly.

The Tucker Low-rank Classification Model for Tensor Data, Junge Li, Florida State University

With the rapid advances of modern technology, tensor data (i.e., multiway array) have been collected in various scientific research and engineering applications. The classification of tensor data is of great interest, where predictive models and algorithms are proposed for predicting a categorical class label for each tensor-valued sample. Aiming to improve interpretability of tensor classification methods, we consider an intuitive and efficient discriminant analysis approach, referred to as the Tucker Low-rank Classification (TLC) model. The TLC model assumes that the between-class mean differences have a low-rank Tucker decomposition, while the covariance matrix is separable. As such, the TLC model greatly reduces the number of parameters by exploiting the tensor structure. We construct a penalized estimator for the TLC model to achieve a sparse Tucker decomposition on the key discriminant analysis parameters and to further improve the parsimony in the final classifier. We establish estimation, variable selection, and prediction consistency for the penalized estimator to confirm that the proposed estimator achieves efficiency gain compared to standard methods. We demonstrate the superior performance of TLC in extensive simulation studies and real data examples.

De-correlated Nearest Shrunken Centroids for Tensor Data, Munwon Yang, Florida State University

The nearest shrunken centroids (NSC) method is an efficient and accurate classifier. However, it is challenged by the recently prevalent tensor data, as it does not model the correlation structure among the tensor predictors. We tackle this challenge by proposing a new distance-based classifier, TDNSC. TDNSC leverages the popular separable covariance structure on tensor data to decorrelate data and allow easy application of NSC afterwards. The theoretical properties and empirical results suggest that TDNSC is a promising method for tensor classification.

Improving Deep Learning Predictions of Regulatory Effects in Genetic Variations through Data Augmentation Using Massively Parallel Reporter Assay Data, Weijia Jin, University of Florida

Background: MPRA (Massively Parallel Reporter Assay) investigates the regulatory effects of genetic variants, typically predicted by Convolutional Neural Network (CNN) models using genetic sequences. However, CNNs demand extensive training datasets, presenting a challenge due to the limited size of MPRA datasets. This study aims to bridge this disparity by developing data augmentation methods.

Methods: We utilized MPRA datasets from six studies, each categorized into 'significant regulatory effect' (positive), 'no significant regulatory effect' (negative), and 'unlabeled' variants. We developed two data augmentation methods: ‘mpraSEM’ employing semi-supervised learning to iteratively train a CNN model using labeled variants and assign pseudo labels to unlabeled variants, and ‘mpraVAE’ utilizing a Conditional Variational Autoencoder architecture to generate pseudo sequences resembling positive or negative sequences by minimizing a weighted combination of three distinct loss functions. Methods were benchmarked against 'naïve augmentation methods', which rely on real sequences around the genetic variants, and several existing machine learning methods for providing whole-genome functional scores.

Results: Our analysis indicates that the models trained with data augmented by 'mpraSEMI' and 'mpraVAE' demonstrated notable improvements in performance metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC), comparing to 'naïve augmentation methods' and 16 previously established methods across all six datasets and two downsampling scenarios.

Discussion: This study introduces two data augmentation methods aimed at enhancing deep learning predictions of regulatory effects, facilitating accurate pre-identification of key regulatory elements. Moreover, these methods offer versatility, as they can be adapted for predictions based on DNA sequences.

Sparse t-Distribution Discriminant Analysis, Xiaoyi Wang, Florida State University

Sparse Linear Discriminant Analysis (LDA) methods are known to yield accurate results in high-dimensional classification problems, and achieve mini-max optimal rate in estimation, prediction and variable selection. However, properties of the existing methods rely on light-tailed assumptions, which is inappropriate with heavy-tailed Big Data containing non-ignorable outliers. Motivated by this, we propose a robust generalization of sparse LDA using t-distribution (TSDA) that remain empirically effective and theoretically optimal with heavy-tail data. We explicitly model the heavy-tail structure by posing t-distribution assumption on features and develop a linear classifier. We then obtain novel one-step mean and covariance matrix estimators that leverage the impact of heavy tails and lead to improved estimation and prediction performance. We rigorously show that TSDA has the same minimax optimal convergence rate as existing sparse LDA methods under weak finite second moments conditions. Further simulation studies and real data analysis prove that TSDA outperforms other competitors when heavy tails are present and remain robust when heavy tails are insignificant.

6:00pm – 6:50pm **Keynote speaker: Ji-Hyun Lee,** University of Florida

Organizer & Chair: Xin (Henry) Zhang

(Longmire 0201) **Statisticians' Role and Leadership in data-driven landscaping**

6:50pm **Dinner Banquet (Longmire Beth Moor Lounge)**

Saturday, March 30 - FLH 0275

8:00am – 9:00am Business meeting of FL chapter

9:00am – 10:00am **Keynote Speaker: Hulin Wu,** UT Health at Houston

Organizer & Chair: Xin (Henry) Zhang

Two Types of Big Data: What Role Can Statisticians Play?

Statistics is at a crossroads and challenged by the emerging discipline of Data Science and AI technologies in the era of Big Data. Originally statisticians were considered as a unique profession to deal with and analyze data, now data scientists and AI technologies are rapidly growing to take over statistician's job. In this talk, I will explore the evolving role of statisticians in the era of Big Data. I divide the Big Data into two distinct types: Type I, which aggregates numerous small datasets through data sharing and curation (e.g., NIH data repositories), and Type II Real World Data, collected from business operations and practices (e.g., Electronic Health Records). I will share my experiences and insights from working with Gene Expression Omnibus (GEO) databases and Electronic Health Records (EHR), highlighting the importance of multidisciplinary collaboration, computational power, and innovative analytical pipelines in extracting meaningful insights from complex Big Data. As the boundaries between statistics, data science, and artificial intelligence continue to blur, our statisticians need to adapt, innovate, and embrace the interdisciplinary nature of our work. By leveraging our unique expertise in handling randomness and uncertainty, we can play a crucial role in the future of data science, contributing to interdisciplinary advancements of sciences.

10:00am – 10:30pm Break

10:30am – 12:00pm **Invited session B – Biostaistics and Bioinformatics** (Organizer & Chair: Zhaotong Lin)

10:30am – 11:00am **Xiaoming Liu**, University of South Florida

MetaRNN: Differentiating Rare Pathogenic and Rare Benign Missense SNVs and InDels Using Deep Learning

Multiple computational approaches have been developed to improve our understanding of genetic variants. However, their ability to identify rare pathogenic variants from rare benign ones is still lacking. Using context annotations and deep learning methods, we present pathogenicity prediction models, MetaRNN and MetaRNN-indel, to help identify and prioritize rare nonsynonymous single nucleotide variants (nsSNVs) and non-frameshift insertion/deletions (nfINDELs). We use independent test sets to demonstrate that these new models outperform state-of-the-art competitors and achieve a more interpretable score distribution. Importantly, prediction scores from both models are comparable, enabling easy adoption of integrated genotype-phenotype association analysis methods. All pre-computed nsSNVs scores are available at http://www.liulab.science/MetaRNN. The stand-alone program is also available at https://github.com/Chang-Li2019/MetaRNN.

11:00am – 11:30am **Guanyu Hu**, UT Health at Houston

Bayesian nonparametric clustering for spatially resolved transcriptomics data

The emergence of spatial molecular profiling technologies has facilitated the acquisition of comprehensive molecular profiles with high resolution while preserving spatial and morphological contexts. However, the dimensionality of molecular profiles in SRT data presents a challenge for existing clustering methods, which often rely on dimension reduction techniques like PCA and UMAP. To our knowledge, there is currently no method that directly models high-dimensional count molecular profile data, and most clustering frameworks assume that the number of clusters is known, which is not the case in real SRT data. To address these issues, we propose a Bayesian nonparametric clustering framework for SRT data (BNPSpace) that directly utilizes count molecular profile data, simultaneously selects the number of clusters, and incorporates spatial information through the Markov Random Field prior. Our framework reduces the dimensionality of molecular data while effectively retaining features relevant to clustering. We evaluate BNPSpace against current clustering frameworks for spatial and non-spatial transcriptomic data using simulations and real SRT data and demonstrate its ability to improve the identification of subpopulations in transcriptional profiles while identifying representative genes for each subpopulation.

11:30am – 12:00pm **Rhonda Bacher,** University of Florida

scLANE: Single-**c**ell **L**inear **A**daptive **N**egative-binomial **E**xpression Testing

Single cell RNA-sequencing (scRNA-seq) has advanced our ability to obtain high-resolution views of dynamic biological processes such as cellular differentiation and disease progression. Many methods have emerged that estimate a cell-level ordering from snapshot scRNA-seq samples by using similarity of gene expression to place cells along a trajectory. With the goal of making biological inferences regarding gene expression across or between trajectories, researchers have typically turned to generalized additive models (GAMs) to capture the complex and nonlinear trends. However, their flexibility comes at a cost to interpretability. To address this, we developed **s**ingle-**c**ell **L**inear **A**daptive **N**egative-binomial **E**xpression (scLANE) testing. Our method balances the need for a nonlinear model to accurately characterize changes in expression while enabling direct biological interpretation. We demonstrate our method's accuracy and ability to draw meaningful comparisons on simulated data and a case-study datasets having tens of thousands of cells and from multiple subjects.

12:00pm – 12:15pm Closing remarks, Student awards ceremony

Organizing Committee:

- Dr. Xin (Henry) Zhang
- Dr. Chao Huang
- Dr. Zhaotong Lin
- Dr. Joshua Loyal
- Dr. Xiulin Xie
- Dr. Jonathan Stewart

Local Organizing Committee:

- Dr. Xin (Henry) Zhang
- Dr. Chao Huang
- Dr. Zhaotong Lin
- Dr. Joshua Loyal
- Dr. Xiulin Xie
- Dr. Jonathan Stewart
- Pamela McGhee
- Zoe Garcia

Student Award Judges: