The Southern Regional Council on Statistics
59th Summer Research Conference

SRCOS
Southern Regional Council
on Statistics

June 3-5, 2024
Clemson University, Clemson, South Carolina

Welcome to the 59th Summer Research Conference of the Southern Regional Council on Statistics! The SRC brings together leading researchers, faculty and students in a dynamic, interactive and learning environment.

We gratefully recognize these 2024 SRCOS SRC Sponsors

NSF

ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

NISS | *The Statistics Community Serving the Nation*

We also thank Dr. Mike Kutner for his generous contribution

Program Schedule
Southern Regional Council on Statistics Summer Research Conference 2024

Clemson University, Clemson, South Carolina

Program Schedule

Sunday, June 2, 2024

5:00 – 7:00 pm Registration

Monday, June 3, 2024

7:30 – 8:15 am Registration

8:15 – 8:30 am Opening Remarks

8:30 – 10:00 am Session I: Statistics and Artificial Intelligence

James (Steve) Marron, Amos Hawley Distinguished Professor, University of North Carolina at Chapel Hill, *Synergies Between 2 Cultures: Statistics and Machine Learning & AI.*

Ashwin Pananjady, Assistant Professor, Georgia Institute of Technology, *Sharply Predicting the Behavior of Complex Iterative Algorithms in Statistical Settings.*

Anirban Bhattacharya, Professor, Texas A&M University, *On the Convergence of Coordinate Ascent Variational Inference.*

10:00 – 10:30 am Coffee Break

10:30 – 12:00 pm Session II: Precision Health

Lu Wang, Professor, University of Michigan, *Statistical Causal Learning Methods for the Optimal Dynamic Decision Rules Leading Toward Personalized Health Care.*

Nikki Freeman, Assistant Professor, Duke University, *Bayesian Machine Learning Strategies for Translational Precision Medicine.*

Zhengling Qi, Assistant Professor, George Washington University, *A Policy Gradient Method for Confounded POMDPs.*

12:10 – 1:30 pm **ASA/Kutner Faculty Poster Session** and **NSF/Harshbarger Student Poster Session**

1:30 – 7:00 pm Discussion, networking and dinner on your own

7:30 – 9:00 pm Session III: Keynote Speaker

Bhramar Mukherjee, John D. Kalbfleisch Distinguished University Professor, University of Michigan, *Unveiling Bias: A Statistician's Quest for Data Equity in Health Research.*

Program Schedule

Tuesday, June 4, 2024

8:30 – 10:00 am Session IV: Experimental Design/Uncertainty Quantification

David Edwards, Professor and Chair, Virginia Commonwealth University, *A Graphical Comparison of Screening Designs using Support Recovery Probabilities.*

Lulu Kang, Associate Professor, University of Massachusetts Amherst, *Optimal Kernel Learning for Gaussian Process Models with High-Dimensional Input.*

Li-Hsiang Lin, Assistant Professor, Georgia State University, *Exact Large-Scaled Inference and Uncertainty Quantification in Varying Coefficient Models with Applications.*

10:00 – 10:30 am Coffee Break

10:30 – 12:00 pm Session V: Environmental/Climate Statistics

Mikyoung Jun, ConocoPhillips Data Science Professor, University of Houston, *Flexible Multivariate Spatiotemporal Hawkes Process Models of Terrorism.*

Ben Seiyon Lee, Assistant Professor, George Mason University, *A Class of Models for Large Zero-inflated Spatial Data*

Staci Hepler, Associate Professor, Wake Forest University, *Spatio-temporal Forecasting for the U.S. Drought Monitor*

12:10 – 1:30 pm Lunch

1:30 – 5:00 pm Discussion, Networking, Free Time

5:00 – 6:00 pm Reception

6:00 – 7:00 pm Banquet – Madren Center Ballroom

7:00 – 7:45 pm Banquet Speaker

Simon Sheather, Dean & Truist Endowed Chair in Data Analytics, University of Kentucky, *The Life and Times of Arnold Stromberg.*

7:45 – 8:30 pm Awards Ceremony and Photos

Program Schedule

Wednesday, June, 5, 2024

8:30 – 10:00 am Session VI: High Dimensional/Highly Structured Data

Ana-Maria Staicu, Professor, North Carolina State University, *Classification of Social Media Users Using Functional Data Analysis.*

Zhengwu Zhang, Assistant Professor, University of North Carolina at Chapel Hill, *Continuous and Atlas-free Analysis of Brain Structural Connectivity.*

Ray Bai, Assistant Professor, University of South Carolina, *Two-step Mixed-type Multivariate Bayesian Sparse Variable Selection with Shrinkage Priors.*

10:00 – 10:15 am Coffee Break

10:15 - 11:30 am Session VII: Statistics Education

Edward Boone, Professor, Virginia Commonwealth University, *Teaching Artificial Intelligence in the Classroom Successes and Pitfalls.*

Megan Mocko, Lecturer, University of Florida, *Communicating about Data across International Boundaries*

David Hitchcock, Associate Professor, University of South Carolina, *A Discussion-based Course on the History of Statistics (With a Little Help From My Friends).*

2024 Keynote Speaker

Dr. Bhramar Mukherjee is John D. Kalbfleisch Distinguished University Professor, Siobán D. Harlow Collegiate Professor of Public Health, and Chair of Biostatistics; Professor of Epidemiology and Global Public Health, UM School of Public Health. She is Associate Director for Quantitative Data Sciences, University of Michigan Rogel Cancer Center as well as Assistant Vice President for Research for Research Data Services Strategy. Bhramar is the founding director of the University of Michigan's summer institute on Big Data (BDSI). Her research interests include statistical methods for analysis of electronic health records, studies of gene-environment interaction, Bayesian methods, and shrinkage estimation, with strong collaborative areas mainly in cancer, cardiovascular diseases, reproductive health, and exposure science. She has co-authored more than 380 peer-reviewed publications in a variety of prestigious academic journals. She is the recipient of many awards for her scholarship, service, and teaching at the University of Michigan and beyond, including a membership with the National Academy of Medicine. Bhramar and her team took an active role in modeling the SARS-CoV-2 virus trajectory in India during the COVID-19 pandemic, which has been covered by major media outlets across the world.

Abstracts for Oral Presentations

(In order of appearance)

Marron, James, University of North Carolina at Chapel Hill
marron@unc.edu

Synergies Between 2 Cultures: Statistics and Machine Learning & AI

Abstract: The close relationships, and also contrasts, between Statistics and Machine Learning (AI) are explored. Both modern ways of thinking and historical background are considered. Important points are illustrated with real data examples.

Pananjady, Ashwin, Georgia Institute of Technology
ashwinpm@gatech.edu

Sharply Predicting the Behavior of Complex Iterative Algorithms in Statistical Settings

Abstract: Iterative algorithms are the workhorses of modern statistical learning and signal processing, and are widely used to fit complex statistical models to data. While the choice of an algorithm and its hyperparameters determines both the speed and fidelity of the learning pipeline, it is common for this choice to be made heuristically, either by expensive trial-and-error or by comparing upper bounds on convergence rates of various candidate algorithms. Motivated by these issues, we develop a principled framework that produces sharp, iterate-by-iterate characterizations of solution quality for complex iterative algorithms on several nonconvex model-fitting problems in statistical settings. Such sharp predictions can provide precise separations between families of algorithms while also revealing nonstandard convergence phenomena. In this talk, I will introduce a general framework to obtain such predictions and showcase it on several canonical high-dimensional statistical models.

Bhattacharya, Anirban, Texas A&M University
anirbanb@stat.tamu.edu

On the Convergence of Coordinate Ascent Variational Inference

Abstract: As a computational alternative to Markov chain Monte Carlo approaches, variational inference (VI) is becoming increasingly popular for approximating intractable posterior distributions in large-scale Bayesian models due to its comparable efficacy and superior efficiency. Several recent works provide theoretical justifications of VI by proving its statistical optimality for parameter estimation under various settings; meanwhile, formal analysis on the algorithmic convergence aspects of VI is still largely lacking. In this talk, we will discuss some recent advances towards studying convergence of the popular coordinate ascent variational inference algorithm. We will present some specific case studies and proceed to develop a general framework for studying such questions.

Wang, Lu, University of Michigan
luwang@umich.edu

Statistical Causal Learning Methods for the Optimal Dynamic Decision Rules Leading Toward Personalized Health Care

Abstracts for Oral Presentations

(In order of appearance)

Abstract: In this talk, we present recent advances and statistical causal learning developments for evaluating Dynamic Treatment Regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. We will first present a tree-based doubly robust reinforcement learning (T-RL) method, which builds a decision tree that maintains the nature of batch-mode reinforcement learning, and then a new Stochastic-Tree Search method called ST-RL for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances even with a large number of covariates. In addition, we consider a common challenge with practical "restrictions" and develop a Restricted Tree-based Reinforcement Learning (RT-RL) method to address this challenge. We illustrate the method using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

Qi, Zhengling, Georgia Washington University
qizhengling@email.gwu.edu
A Policy Gradient Method for Confounded POMDPs

Abstract: In this paper, we propose a policy gradient method for confounded partially observable Markov decision processes (POMDPs) with continuous state and observation spaces in the offline setting. We first establish a novel identification result to non-parametrically estimate any history-dependent policy gradient under POMDPs using the offline data. The identification enables us to solve a sequence of conditional moment restrictions and adopt the min-max learning procedure with general function approximation for estimating the policy gradient. We then provide a finite-sample non-asymptotic bound for estimating the gradient uniformly over a pre-specified policy class in terms of the sample size, length of horizon, concentratability coefficient and the measure of ill-posedness in solving the conditional moment restrictions. Lastly, by deploying the proposed gradient estimation in the gradient ascent algorithm, we show the global convergence of the proposed algorithm in finding the history-dependent optimal policy under some technical conditions. To the best of our knowledge, this is the first work studying the policy gradient method for POMDPs under the offline setting.

Freeman, Nikki, Duke University
nikki.freeman@duke.edu
Bayesian Machine Learning Strategies for Translational Precision Medicine

Abstract: Precision medicine is a promising framework for generating evidence to improve health and health care. Yet, a gap persists between the ever-growing number of statistical precision medicine strategies for evidence generation and implementation in real-world clinical settings, and the strategies for closing this gap will likely be context-dependent. In this talk, we will provide a friendly introduction to statistical precision medicine and what, to this point, have been common precision medicine tasks, evaluating treatment rules and learning optimal treatment rules. We will then consider two Bayesian approaches that attempt to bridge the aforementioned gap—a strategy that uses a Gaussian process surrogate for the value function to quickly characterize classes of optimal dynamic treatment regimes and a Bayesian approach to outcome weighted learning that enables uncertainty quantification of individual decisions. Throughout, we will emphasize how these findings aid the translation of precision medicine evidence into clinical contexts.

Mukherjee, Bhramar, University of Michigan
bhramar@umich.edu
Unveiling Bias: A Statistician's Quest for Data Equity in Health Research

Abstract: Despite numerous proposed strategies to enhance diversity in scientific research, a significant portion of the world's health research data continues to be derived from populations with historical privileges. In this presentation, I will delve into the crucial concept of data equity and highlight the obvious: algorithms developed on exclusionary datasets can yield erroneous conclusions and exacerbate health disparities. However, while we strive for the ideal scenario of globally representative and extensive datasets, statisticians play a pivotal role in mitigating selection bias and handling missing data—an expertise that few other quantitative disciplines possess. Drawing from my own journey as a statistician, I will showcase instances where timely statistical analysis with imperfect data resulted in enhanced inference and influenced policy outcomes. Utilizing examples from analyses of COVID-19 and biobanks linked with electronic health records, I will demonstrate how statistical methodologies can navigate the complexities of real-world data to inform decision-making and drive positive change. In conclusion, I urge statisticians to lead efforts in curating, and collecting new data, launching their own studies while also spearheading innovative scientific inquiries. It is time for statisticians to step out from the fringes of design and analysis and assert their independent leadership in shaping the trajectory of research and policymaking driven by data.

Edwards, David, Virginia Commonwealth University
dedwards7@vcu.edu
A Graphical Comparison of Screening Designs using Support Recovery Probabilities

Abstract: A screening experiment attempts to identify a subset of important effects using a relatively small number of experimental runs. Given the limited run size and a large number of possible effects, penalized regression is a popular tool used to analyze screening designs. In particular, an automated implementation of the Gauss-Dantzig selector has been widely recommended to compare screening design construction methods. In this talk, we illustrate potential reproducibility issues that arise when comparing two-level screening designs via simulation, and recommend a graphical method, based on screening probabilities, which compares designs by evaluating them along the penalized regression solution path. This method can be implemented using simulation, or, in the case of lasso, by using exact local lasso sign recovery probabilities. Our approach circumvents the need to specify tuning parameters associated with regularization methods, leading to more reliable design comparisons.

Kang, Lulu, University of Massachusetts, Amherst
lulukang@umass.edu
Optimal Kernel Learning for Gaussian Process Models with High-Dimensional Input

Abstract: Many computer simulation models in engineering and scientific domains involve a large number of input variables, which can result in high computational cost and low prediction accuracy for the Gaussian process (GP) regression model. However, some simulation models may only be significantly influenced by a small subset of the input variables, referred to as the “active variables”. Identifying these active variables can help researchers overcome the two limitations of the GP model and gain a better understanding of the simulated system. To achieve this goal, we propose an approximation of the covariance function of the original GP model involving all the input variables. The approximation is through a convex combination of kernel functions whose input variables are low-dimensional subsets of the complete input variables. To determine the optimal approximation, we develop an iterative algorithm based on the Fedorov-Wynn algorithm from the optimal design literature. We also incorporate the effect heredity principle while selecting the active input variables, which ensures sparsity. Through several examples, we have shown the proposed method outperforms some alternative approaches in correctly identifying the active input variables.

Lin, Li-Hsiang, Georgia State University
lhlin@gsu.edu
Exact Large-Scaled Inference and Uncertainty Quantification in Varying Coefficient Models with Applications

Abstract: This study is motivated by the need to analyze a large-scale household migration dataset from coastal Louisiana. The household migration patterns are known to be related to several social/economic factors whose effects change dynamically according to time and spatial information, motivating the use of varying coefficient models. However, the sample size is too large to be processed into an analysis software, making the conventional inference methods impractical. To address this challenge, we divided the data into non-overlapping subdatasets, each small enough to be managed by the analysis software. We then identified the sufficient statistics within each subdataset.

This approach significantly reduces computational complexity, allowing us to quantify regression effects and their uncertainty without processing the entire dataset at once. Remarkably, our proposed inference method does not rely on any approximations, ensuring our estimates are the same as they would be if the entire dataset could be used. Intensive numerical studies and the analysis of the motivation dataset will be presented. Also, we will discuss the extensions of our method to other models and applications.

Jun, Mikyoung, University of Houston
mjun@central.uh.edu
Flexible Multivariate Spatiotemporal Hawkes Process Models of Terrorism

Abstract: We develop flexible multivariate spatiotemporal Hawkes process models to analyze patterns of terrorism. Previous applications of point process methods to political violence data mainly utilize temporal Hawkes process models, neglecting spatial variation in these attack patterns. This limits what can be learned from these models, as any effective counter-terrorism strategy requires knowledge on both when and where attacks are likely to occur. Even the existing work on spatiotemporal Hawkes processes imposes restrictions on the triggering function that are not well-suited for terrorism data. Therefore, we generalize the structure of the spatiotemporal triggering function considerably, allowing for nonseparability, nonstationarity, and cross-triggering (across multiple terror groups). To demonstrate the utility of our models, we analyze two samples of real-world terrorism data: Afghanistan (2002–2013) as a univariate analysis and Nigeria (2009–2017) as a bivariate analysis. Jointly, these two studies demonstrate that our generalized models outperform standard Hawkes process models, besting widely-used alternatives in overall model fit and revealing spatiotemporal patterns that are, by construction, masked in these models (e.g., increasing dispersion in cross-triggering over time). This is joint work with Scott Cook at Department of Political Science, Bush School of Government and Public Service, Texas A&M University.

Lee, Ben Seiyon, George Mason University
slee287@gmu.edu
A Class of Models for Large Zero-inflated Spatial Data

Abstract: Spatially correlated data with an excess of zeros, usually referred to as zero-inflated spatial data, arise in many disciplines. Examples include count data, for instance, abundance (or lack thereof) of animal species and disease counts, as well as semi-continuous data like observed precipitation. Spatial two-part models are a flexible class of models for such data. Fitting two-part models can be computationally expensive for large data due to high-dimensional dependent latent variables, costly matrix operations, and slow mixing Markov chains. We describe a flexible, computationally efficient approach for modeling large zero-inflated spatial data using the projection-based intrinsic conditional autoregression (PICAR) framework. We study our approach, which we call PICAR-Z, through extensive simulation studies and two environmental data sets. Our results suggest that PICAR-Z provides accurate predictions while remaining computationally efficient. An important goal of our work is to allow researchers who are not experts in computation to easily build computationally efficient extensions to zero-inflated spatial models; this also allows for a more thorough exploration of modeling choices in two-part models than was previously possible. We show that PICAR-Z is easy to implement and extend in popular probabilistic programming languages such as nimble and stan.

Hepler, Staci, Wake Forest University
heplersa@wfu.edu
Spatio-temporal Forecasting for the U.S. Drought Monitor

Abstract: Drought is a natural hazard that has a significant impact on sustainability. Drought leads to water scarcity, which impacts health, agriculture, and the economy. Better understanding of causes and impacts of drought can result in better management and mitigate the effects. The US Drought Monitor is the leading drought monitoring tool in the United States. It records weekly drought conditions as geo-referenced polygons showing one of six ordered levels. These levels are determined by a mixture of quantitative environmental measurements and local expert opinion. At present, forecasts of the Drought Monitor only convey the expected direction of drought development and do not communicate uncertainty. We develop a Bayesian spatio-temporal ordinal model for modeling and projecting drought conditions. The modeling framework specified is flexible, scalable, and interpretable. A two-stage approach to model fitting permits parallel computing and alleviates computational expense associated with large space-time data. We use our model to produce future forecasts of actual drought levels – rather than only the direction of drought development as is done with existing drought forecast tools.

Sheather, Simon, University of Kentucky
simon.sheather@uky.edu
The Life and Times of Arnold Stromberg

Abstract: This talk will explore the remarkable life and career of Arnold Stromberg.

Staicu, Ana-Maria, North Carolina State University
ana-maria_staicu@ncsu.edu
Classification of Social Media Users Using Functional Data Analysis

Abstract: Technological advancement has made possible the collection of data from social media platforms at unprecedented speed and volume. Current methods for analyzing such data either lack interpretability, are computationally intense or require a rigid data regimen. In this work, we propose a flexible statistical framework for the analysis of high-resolution data arising from social media applications. We focus on the posting behavior of social media users and consider functional data-based methods to extract relevant information of a user's posting behavior and ultimately identify the type of account (malicious or genuine). We illustrate the methods numerically and on a motivating Twitter data set. The developed methods are applicable to other social media data, such as Facebook, Instagram, Reddit, or TikT0k, or any form of digital interaction where user's posting behavior is a key feature of their type.

Zhang, Zhengwu, University of North Carolina at Chapel Hill
zhengwu_zhang@unc.edu

Continuous and Atlas-free Analysis of Brain Structural Connectivity

Abstract: Brain structural networks are often represented as discrete adjacency matrices with elements summarizing the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined a-priori using a brain atlas. The choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This work introduces an atlas-free framework that overcomes these issues by modeling brain connectivity using smooth random functions. In particular, we assume that the observed pattern of white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain. To facilitate statistical analysis of these high dimensional functional data objects, we develop a novel algorithm to construct a data-driven reduced-rank function space that offers a desirable trade-off between computational complexity and flexibility. Using real data from the Human Connectome Project, we show that our method outperforms state-of-the-art approaches that use the traditional atlas-based structural connectivity representation on a variety of connectivity analysis tasks. We further demonstrate how our method can be used to detect localized regions and connectivity patterns associated with group differences.

Bai, Ray, University of South Carolina
rbai@mailbox.sc.edu

Two-step Mixed-type Multivariate Bayesian Sparse Variable Selection with Shrinkage Priors

Abstract: We introduce a Bayesian framework for mixed-type multivariate regression using continuous shrinkage priors. Our framework enables joint analysis of mixed continuous and discrete outcomes and facilitates variable selection from the p covariates. Theoretical studies of Bayesian mixed-type multivariate response models have not been conducted previously and require more intricate arguments than the corresponding theory for univariate response models due to the correlations between the responses. In this talk, we investigate necessary and sufficient conditions for posterior contraction of our method when p grows faster than sample size n. The existing literature on Bayesian high-dimensional asymptotics has focused only on cases where p grows subexponentially with n. In contrast, we study the asymptotic regime where p is allowed to grow exponentially terms of n. We develop a novel two-step approach for variable selection which possesses the sure screening property and provably achieves posterior contraction even under exponential growth of p. We demonstrate the utility of our method through simulation studies and an application to a cancer genomics dataset where n=174 and p=9183.

Boone, Edward, Virginia Commonwealth University
elboone@vcu.edu

Teaching Artificial Intelligence in the Classroom Successes and Pitfalls

Abstract: As Artificial Intelligence becomes more a part of our everyday lives, there is a need for people who understand how these techniques work. In two of my classes I taught AI from an Artificial Neural Network standpoint in both R and Python. The course where R was used was an introductory statistical programming course and several R packages were used. The other course was a second course in statistical computing where Python was introduced. In both of these courses the overall idea of an Artificial Neural Network was presented with a few examples. The course with Python considered classification problems, network structure problems, function fitting, convolutional neural networks, large language models for sentiment analysis and a simple version of chat GPT. I will present many of the issues one is likely to face when teaching these topics in the course as well as the successes that students have on the topic.

Mocko, Megan, University of Florida
Megan.Mocko@warrington.ufl.edu

Communicating about Data across International Boundaries

Abstract: With the world becoming increasingly interconnected, preparing students to talk about data and data ethics with individuals different from themselves will become an increasingly important skill. In virtual exchange, instructors from different universities, usually in different countries, work together to create a common activity for students to work on together. In this session, I will discuss several different versions of virtual exchange experiences that I have conducted with partners in Ecuador, Mexico, Colombia, Scotland and the UK. I will also share research from these experiences and future questions.

Hitchcock, David, University of South Carolina
hitchcock@stat.sc.edu

A Discussion-based Course on the History of Statistics (With a Little Help From My Friends)

Abstract: A special-topics undergraduate course about the history of statistics which was taught in Spring 2023 at the University of South Carolina is described. We review other similar courses (past and current) and explain the discussion-based nature of this course. The conception and planning of the course are detailed, and the unique experiences (activities, guest speakers, presentations, etc.) are described. The course emphasized substantial amounts of independent reading outside of class and lively discussions during class. Topics covered in the class include the early development of probability, the normal distribution, and the central limit theorem; the development of modern statistical science by British statisticians; the rise of formal mathematical statistics; and increasing specialization and modern computational and data-analytic advances. An assessment of the course's effectiveness based on qualitative student survey data is given. Students were highly complimentary of the course, praising the in-class discussion format, the benefits of doing the outside readings, the invited guest speakers, and the in-class activities. There were occasional comments that the amount of required reading was excessive. Based on this, suggestions for future offerings of the course are presented, including developing a more carefully curated set of readings.

Abstracts for ASA/Kutner Faculty Posters

1 *Estimation of Area Deprivation Index Using Spatiotemporal Change of Support*

Hossein Moradi Rekabdarkolaee, South Dakota State University

Abstract: A common issue in spatial and temporal statistical analysis occurs when we want to make inferences about a variable, but the spatial or temporal support of the observed data does not match the desired support. The process of transforming data to the desired support is referred to as change of support (COS). The traditional approach for performing spatial-only COS is to estimate values based on areal proportions. This method is easy to implement but works only for spatial COS and does not provide measures of uncertainty. Furthermore, there is no reliable way to evaluate the performance of this method. In this project, we employed a spatiotemporal change of support (STCOS) model which allows for both spatial and temporal COS on Gaussian data. The model uses a ”bottom-up” approach and a Bayesian hierarchical model framework. This methodology can provide model-based estimates, predictions, and associated measures of uncertainty. We present a case study using the national Area Deprivation Index (ADI) rankings for the state of South Dakota. ADI rankings are used to identify areas of socioeconomic disadvantage at the census block group level. The STCOS model is demonstrated by estimating the ADI rankings of ZIP Code Tabulation Areas in South Dakota.

2 *Trigonometry-transformation Based Correlation Coefficient with an Application to Sufficient Variable Selection*

Pei Wang, Miami University

Abstract: The technique of variable selection has gained widespread popularity for reducing data size, particularly in the context of large p small n datasets. In this study, we introduce a novel criterion based on the correlation coefficient derived from trigonometry transformations. This innovative criterion serves as a metric for assessing the relationship between the response and each predictor. When integrated into a two-step selection procedure, it becomes a valuable tool for variable selection. Notably, this approach is model-free, providing robustness against model mis-specification. We establish the asymptotic and sure selection properties, and the effectiveness of the proposed method is demonstrated through extensive numerical studies and real data analysis.

3 *A Projection-based Test for Drug Discovery using Multivariate Longitudinal Data*

Salil Koner, Duke University

Abstract: Contemporary longitudinal studies gather various outcomes as key indicators to comprehend the intricate dynamics of diseases. Particularly in clinical trials, understanding the collective variation among the multifaceted responses is pivotal in evaluating differences between multiple groups, rather than solely relying on a single outcome. We devise a projection-based two-sample significance test to pinpoint population-level disparities among the multivariate profiles observed in a sparse longitudinal setup. This approach, founded on widely embraced multivariate functional principal component analysis, diminishes dimensionality of infinite-dimensional multi-modal functions while retaining the dynamic correlation among components. The test is applicable across a broad range of (non-stationary) covariance structures of the response, detecting significant group disparities through a single p-value, thereby sidestepping the need for adjusting multiple p-values resulting from comparing means in individual components separately. Finite-sample numerical analyses validate the test's ability to maintain type-I error and its potency in detecting significant group differences compared to leading testing techniques. The test is implemented in two notable longitudinal studies involving Alzheimer's and Parkinson's disease patients, namely, TOMMORROW study examining individuals at high risk of mild cognitive impairment to discern variations in cognitive test scores between pioglitazone and placebo groups, and Azillect study assessing rasagiline's efficacy as a potential treatment to decelerate Parkinson's disease progression.

4 *Robust Estimation and Variable Selection for Linear Mixed Models*

Olivia Atutey, University of South Alabama

Abstract: A BIC-type robust variable selection procedure is proposed to select and estimate fixed effects in linear mixed models simultaneously. Robustness in the presence of contaminated, heavy-tailed symmetric data distributions and/or outliers, and sparse estimation of fixed effects are achieved by adapting Jaeckel's dispersion function based on the absolute values of the ranks of the residuals for a given choice of score function with hyperbolic tangent penalty function. The finite sample behavior of the proposed rank-based approach is evaluated via extensive simulation studies under contaminated and heavy-tailed symmetric model error distributions. In addition, a dataset on Medicare spending on a cohort of colon cancer patients is used to illustrate the proposed methodology further.

5 *Exploring the Opioid Syndemic in North Carolina: A Novel Approach to Modeling and Identifying Factors*

Eva Murphy, Wake Forest University

Abstract: The opioid epidemic is a significant public health challenge both nationally and in North Carolina, yet limited data restricts our understanding of its complexity. Recognizing the synergistic interaction among various epidemics in forming the opioid syndemic provides an alternative perspective to model opioids. Using county-level data spanning 2017 to 2021, consisting of illicit opioid overdose deaths, emergency department visits related to drug overdose, treatment counts for opioid use disorder, patients receiving prescriptions for buprenorphine, and newly diagnosed cases of acute and chronic HCV and HIV, we employ a Bayesian dynamic spatial factor modeling to capture the intricate dynamics within these epidemics. In following this framework, we depart from the conventional lower triangular structure in the loadings matrix, leading to the common identifiability issue in factor modeling. To address this challenge, we propose a novel approach that involves the LQ decomposition on the loadings matrix and allows us to estimate the loadings and factors uniquely. Using a Markov chain Monte Carlo algorithm, we estimate four latent factors that characterize variations across all and subsets of the outcomes in space and time. These factors help us to understand the impact of the opioid epidemic, including its burden, harm, opioid disorder, and its aspect of infectious diseases, along with their spatial distributions and evolution over time.

6 *Surrogate Methods for Analyzing Partial Associations in Mixed Data: Applications to Well-being Survey Analysis*

Zhaohu (Jonathan) Fan, Georgia Institute of Technology

Abstract: This paper is motivated by the analysis of a survey study focusing on college student well-being before and after the COVID-19 pandemic outbreak. A statistical challenge in well-being studies lies in the multidimensionality of outcome variables, recorded in various scales such as continuous, binary, or ordinal. The presence of mixed data complicates the examination of their relationships when adjusting for important covariates. To address this challenge, we propose a unifying framework for studying partial association between mixed data. We achieve this by defining a unified residual using the surrogate method. The idea is to map the residual randomness to a consistent continuous scale, regardless of the original scales of outcome variables. This framework applies to parametric or semiparametric models for covariate adjustments. We validate the use of such residuals for assessing partial association, introducing a measure that generalizes classical Kendall's tau to capture both partial and marginal associations. Moreover, our development advances the theory of the surrogate method by demonstrating its applicability without requiring outcome variables to have a latent variable structure. In the analysis of the college student well-being survey, our proposed method unveils the contingency of relationships between multidimensional well-being measures and micro personal risk factors (e.g., physical health, loneliness, and accommodation) as well as the macro disruption caused by COVID-19.

7 *Testing for Marginal Covariate Effect When the Subgroup Size Induced by the Covariate is Informative*

Samuel Anyaso-Samuel, National Cancer Institute

Abstract: Informative cluster size (ICS) typically introduces bias in cluster-correlated data analyses. We study a complex form of informativeness where the number of observations corresponding to latent levels of a unit-level continuous covariate within a cluster is associated with the response variable. This type of informativeness has not

been explored in prior research. We present a novel test statistic designed to evaluate the effect of the covariate while accounting for informativeness. The covariate induces a continuum of latent subgroups within the clusters, and our test statistic is formulated by aggregating values from an established statistic that accounts for informative subgroup sizes when comparing group-specific marginal distributions. Through simulations, we compare our test with four traditional methods commonly employed in cluster-correlated data analyses. Only our test maintains the size across all data-generating scenarios with informativeness. We illustrate the proposed method to test for marginal associations in periodontal data with this distinctive form of informativeness.

8 *Changepoint Detection in Categorical Time Series with Application to Daily Sky-cloudiness Conditions*

Mo Li, University of Louisiana at Lafayette

Abstract: Changepoint analysis of non-stationary ordinal categorical time series is always of interest in climate studies. For instance, the sky-cloudiness condition in Canada is observed hourly in terms of tenths of the sky dome covered by clouds, and reported by five different categories, sky clear (0), few (1/10-3/10), scattered (4/10-5/10), broken (6/10-9/10), and overcast (10/10), hence having 5 ordinal categories. The cloud cover data often contain changepoints and exhibit temporal trends, seasonality, and serial correlation in nature. To properly take into account these features, a likelihood ratio-type statistic is proposed in this talk to test for a single changepoint in ordinal categorical time series using a marginalized transition model. This model allows for likelihood-based inference, and the series dependence is specified via a first-order Markov chain. An application of our method is illustrated using the daily sky-cloudiness conditions at Fort St. John Airport in Canada, and a stochastic optimization algorithm is adapted to reduce the computation time.

9 *RNA-seq Differential Expression Analysis Accounting for Hidden Factors and Variable Selection*

Yet Nguyen, Old Dominion University

Abstract: In RNA-seq data analysis, a primary objective is the identification of differentially expressed genes, which are genes that exhibit varying expression levels across different conditions of interest. It is widely known that hidden factors, such as batch effects, can substantially influence the differential expression analysis. Furthermore, apart from the primary factor of interest and unforeseen artifacts, an RNA-seq experiment typically contains multiple measured covariates, some of which may significantly affect gene expression levels, while others may not. Existing methods either address the covariate selection or the unknown artifacts separately. In this study, we will investigate several strategies for dealing with both covariate selection and hidden factors via simulation using a real RNA-seq dataset.

Abstracts for ASA/Kutner Faculty Posters

10 *Localized Transfer Learning for a Non-stationary Spatial Model*

Wenlong Gong, University of Houston - Downtown

Abstract: Ambient air pollution measurements from regulatory monitoring networks are routinely used to support epidemiologic studies and environmental policy decision making. However, regulatory monitors are spatially sparse and preferentially located in areas with large populations. Numerical model output can be leveraged into the inference and prediction of air pollution data combining with measurements from monitors. Nonstationary covariance functions allow the model to adapt to spatial surfaces whose variability changes with location like air pollution data. In the paper, we employ localized covariance parameters learned from the numerical output model to knit together into a global nonstationary covariance, and use this nonstationary covariance in a fully Bayesian model in which the unknown spatial process has a Gaussian process prior distribution. We model the nonstationary structure in a computationally efficient way to make the Bayesian model scalable.

11 *Computationally Scalable Bayesian SPDE Modeling for Censored Spatial Responses*

Suman Majumder, University of Missouri

Abstract: Observations of groundwater pollutants, such as arsenic or Perfluorooctane sulfonate (PFOS), are riddled with left censoring. These measurements have impact on the health and lifestyle of the populace. Left censoring of these spatially correlated observations are usually addressed by applying Gaussian processes (GPs), which have theoretical advantages. However, this comes with a cubic computational complexity, which is impractical for large datasets. Additionally, a sizable proportion of the data being left-censored creates further bottlenecks, since the likelihood computation now involves an intractable high-dimensional integral of the multivariate Gaussian density. In this article, we tackle these two problems simultaneously by approximating the GP with a Gaussian Markov random field (GMRF) approach that exploits an explicit link between a GP with Matern correlation function and a GMRF using stochastic partial differential equations (SPDEs). We introduce a GMRF-based measurement error into the model, which alleviates the likelihood computation for the censored data, drastically improving the speed of the model while maintaining admirable accuracy. Our approach demonstrates robustness and substantial computational scalability, compared to state-of-the-art methods for censored spatial responses across various simulation settings. Finally, the fit of this fully Bayesian model to the concentration of PFOS in groundwater available at 24,959 sites across California, where 46.62% responses are censored, produces prediction surface and uncertainty quantification in real time, thereby substantiating the applicability and scalability of the proposed method. Code for implementation is made available via GitHub.

Abstracts for ASA/Kutner Faculty Posters

12 *Gradient Boosting for Group Testing Data*

Erica Porter, Clemson University

Abstract: When screening a population for a disease, it is often necessary or preferable to group or pool individual specimens and perform testing on the pools. This approach is faster and lower cost than testing each individual specimen. It is often of interest to model the true disease status as a function of one or more covariates available for the individuals (e.g. demographic information). We propose a gradient boosting method to build models for group testing data using individual-level covariates. We first develop gradient boosting for the case of masterpool testing, where testing is performed only on the pools with no follow-up testing, and then we describe gradient boosting that can be used with any testing protocol. Upon deriving the gradient for group-testing data, we demonstrate that our approach can be used to build models from a number of weak learners, including regression trees, kernel smoothing, and splines. For each of these weak learners, we develop a cross-validation approach to select appropriate values for applicable tuning parameters. Since the true individual disease statuses are not observed, typical cross-validation metrics such as mean squared error (MSE) cannot be applied. Instead, we calculate the log-likelihood value based on the observed group testing data and choose the tuning parameters that produce the largest log-likelihood value after gradient boosting across several folds. Our gradient boosting approach can be used to model both linear and nonlinear relationships between the disease status and covariates, and can easily be adapted to accommodate different types of weak learners. We demonstrate our method using a data set where group testing was used to screen for chlamydia in Iowa.

13 *Ensemble Learning Model With Effective Feature Selection For Accurately Predicting Failure to Rescue After Coronary Artery Bypass Grafting*

Rameshbabu Manyam, Emory University

Abstract: Failure to rescue (FTR) is defined as a 30-day mortality after a surgical complication and is widely identified as a quality metric in cardiac surgery. Current FTR predictive models use statistical methods that assume linear associations among patient's clinicopathological characteristics. We developed a new ensemble learning (EL) model to handle nonlinear relationships and predict risk factors for FTR after coronary artery bypass grafting (CABG) accurately. We analyzed electronic medical records (EMR) data of adult patients who underwent isolated CABG at a multi-hospital academic health system from 2015 to 2022. The study variables include the patient demographics (e.g., age, gender, race), preoperative comorbidities (e.g. diabetes, hypertension), perioperative vitals, laboratory values, and medications (e.g., body mass index, creatinine, hemoglobin). The primary outcome was FTR. The study evaluated 49 potential risk factors for FTR, using eXtreme gradient Boosting (XGBoost), Recursive Feature Elimination (RFE) with

cross-validation (CV) machine learning (ML) techniques to determine the optimal subset of features with the highest predictive value. Five EL models (based on XgBoost, Random Forest, Support Vector Machine, Neural Networks and Logistic Regression methods) were built on the training data using 10-fold CV and validated on the test data of patients. The relative importance of the risk factors, area under the receiver operating characteristics (AUC), calibration and Brier score were used to assess and quantify the performance of EL algorithms. Preliminary results demonstrated the XGBoost model with good discrimination in identifying patients at high-risk of FTR. The proposed framework can serve as a proof-of-concept for creating intuitive and user-friendly dashboards to provide real time clinical decision support to at-risk patients. EM models can address the issues of multiple and correlated predictors, non-linear associations and interactions between the predictors and outcome, and process massive amounts of EMR data faster than the traditional methods. Further research on larger patient population, more granular data and external validation is necessary to confirm the prediction ability and clinical utility with respect to FTR risk-stratification of patients undergoing CABG.

14 *Bayesian Shrinkage Priors: Harnessing Statistical Efficiency for Real World Applications*

Arinjita Bhattacharyya, Merck

Abstract: We will survey the recent developments of global-local shrinkage priors that have achieved state-of-the-art in high-dimensional data analysis with sparsity and other structural constraints. In the first half, we will provide a brief overview of the history, usage, computational challenges, and optimality guarantees for these priors, covering sequence models, regression, and graphical models. In the second half, we present a new methodology for testing differential abundance in matched-pair count data with zero-inflation or quasi-sparsity. We provide support for excellent small-sample performance with numerical and real-data results and end with scopes for future directions.

15 *Predicting Supply of Long-term Care Insurance with Survival Neural-Network Models and Enterprise Risk Management.*

Sebastain Awondo, University of Alabama

Abstract: Despite the growing need for private insurers to increase their market shares of Long-term Care (LTC), the number of insurers offering LTC insurance (LTCI) coverage has decreased from its peak of over 100 in 2004 to about a dozen in 2020 due to high claim exposure and insolvency. This paper employs neural networks and an enterprise risk management framework to predict insolvency risk using firm-level accounting and financial time-series data, allowing for early interventions by regulators to prevent further deterioration and the demise of LTCIs. An enterprise risk management (ERM) framework seeks to identify, assess, monitor, and proactively manage a portfolio of risks associated

with their business operations. To this end, our empirical approach dwells on a holistic risk management approach to LTCI by incorporating determinants of risk profiles associated with underwriting, investing, credit, and the financial business operations of LTCI to train and validate the predictive model of LTCI insolvency. More specifically, our time-to-event neural network models include as predictors earned premium, incurred claims, loss ratio, number of new policies issued, policyholder persistency rate, policy reserves, population over 65 years, GDP per capita, inflation, interest rate on the 5 and 30-year Treasury, and 63 policy variations over two decades to train, validate, and predict the supply of discontinuation of LTCI.

16 *Time-varying Correlations between JSE.JO Stock Market and its Partners using Symmetric DCC Models.*

Bernard Omolo, University of South Carolina-Upstate

Abstract: The extent of correlation or co-movement among the returns of developed and/or emerging stock markets remains pivotal for efficiently diversifying global portfolios. This correlation is prone to variation over time as a consequence of escalating economic interdependence fostered by international trade and financial markets. In this study, we analyze the time-varying correlation and co-movement between the JSE.JO stock market of South Africa and its developed and developing stock markets. The Dynamic Conditional Correlation - Exponential Generalized Autoregressive Conditional Heteroscedasticity (DCC-EGARCH) methodology is employed with different multivariate distributions to explore the time-varying correlation and volatilities between the JSE.JO stock market and its partners. Based on the conditional correlation results, the JSE.JO stock market is integrated and co-moves with its partners, and the conditional correlation for all markets exhibits time-variant behavior. The conditional volatility results show that the JSE.JO stock market behaves differently from other markets, especially after 2015, indicating a positive sign for investors to diversify between JSE.JO and its partners. The highest value of conditional volatility for markets was in 2020 during the COVID-19 pandemic, representing the riskiest period that investors should avoid due to the lack of diversification opportunities during crises

17 *A Display Portal includes the Tabular Information and Interactive Maps for the IDS Dataset and Annual Fire Dataset of US Forest Service*

Omid Khormali, University of Evansville

Abstract: The health of treed areas, including forests, woodlands, and other vegetated landscapes, is critical. Every year, the Annual Insect and Disease Survey (IDS) dataset collects information on the health status of treed areas impacted by insects and diseases, utilizing both aerial and ground surveys. This study analyzes a portal created by my students in Data Analytics ChangeLab at the University of Evansville. The portal depicts

the tabular information and displays the interactive map of the IDS dataset nationally by year from 2000-2021. It also displays the interactive map of the Annual Fire dataset nationally by year from 1995-2022.

Abstracts for NSF/Harshbarger Student Posters

1 *On Exact Bayesian Credible Sets for Classification and Pattern Recognition*

Song Chaegeun, The Pennsylvania State University

Abstract: The current definition of a Bayesian credible set cannot, in general, achieve an arbitrarily preassigned credible level. This drawback is particularly acute for classification problems, where there are only a finite number of achievable credible levels. As a result, there is as of today no general way to construct an exact credible set for classification. In this paper, we introduce a generalized credible set that can achieve any preassigned credible level. The key insight is a simple connection between the Bayesian highest posterior density credible set and the Neyman-Pearson lemma, which, as far as we know, hasn't been noticed before. Using this connection, we introduce a randomized decision rule to fill the gaps among the discrete credible levels. Accompanying this methodology, we also develop the Steering Wheel Plot to represent the credible set, which is useful in visualizing the uncertainty in classification. By developing the exact credible set for discrete parameters, we make the theory of Bayesian inference more complete.

2 *Optimal Decision of Cure Process: Constrained Bayesian Optimization with Gaussian Process Models*

Yezhuo Li, Clemson University

Abstract: The cure process, a prevalent technique for manufacturing structural components, necessitates the approximation of resultant product deformations. Existing methodologies for deformation estimation are either prohibitively expensive or inherently restricted. To mitigate these challenges, this study introduces a constrained Bayesian Optimization (cBO) strategy, leveraging Gaussian Process (GP) models as surrogate functions. Bayesian Optimization is renowned for its efficacy in navigating complex or black-box functions, while GP models excel in prediction and uncertainty quantification. Through the implementation of our proposed method, we efficiently achieved optimal outcomes at a reduced cost, demonstrating the approach's effectiveness.

3 *Exploring Spatiotemporal Trends in Air Pollutants*

Sukanya Bhattacharyya, North Carolina State University

Abstract: Concern regarding climate change and its influential impact on humanity is the talk of the hour. Air pollutant levels in air are constantly monitored, and we use the United States Environmental Protection Agency's available resources to access the distribution of particular pollutants for a given number of sites, over the years. Various spatial locations have their spatially dependent pollutant's quantile functions which varies with time. Using an approach of simultaneously modelling the quantiles, our aim is to reduce the computational complexity than the existing methodologies. We use a quantile regression method that uses functional principal components to reduce the dimensions

over space and quantile levels while testing for trends in air pollution data over the last 20 years. Extensive comparison among the existing methods in literature is demonstrated.

4 *Computing Bayes factor using the Cross-Entropy method*

Vy Ong, Augusta University

Abstract: The Bayes factor is an alternative to the p-value and is commonly employed for statistical model evaluation. However, calculating the Bayes factor can be formidable when dealing with intractable associated marginal likelihoods. In our research, we have devised a novel method for computing the Bayes factor. This method is built upon the Cross-Entropy technique introduced by Chan and Eisenstat (2012), which combines Markov Chain Monte Carlo (MCMC) with the importance sampling technique to determine the optimal parameters. We illustrate the effectiveness of our novel method through two examples where the marginal likelihood of the first one does have a closed form and that of the second one does not. In these examples, we test whether the mean of a normal distribution is zero or not, with the mean following a normal distribution prior and the precision a Gamma distribution prior. The two priors are dependent in the first example and independent in the second example. We test our method on a real data set and compare it with the existing methods as well.

5 *An Analysis of the Decomposition of Distances Between Covariance Matrices of Protein Binding Sites*

Thambawita Maddumage Sajith Priyankara, Texas Tech University

Abstract: Protein-ligand interactions perform pivotal roles in biological functions, with ligand binding sites (LBSs) on proteins as critical regions for such interactions. Identifying and understanding these sites are valuable for advancement in drug discovery and molecular biology. This study leverages the Covariance of Distances to Principle Axes (CPDA) method that utilizes covariance matrices to encode the structural information of LBSs to analyze a decomposition of these covariance matrices to better understand the performance of CDPA. By examining the Kahraman dataset, which contains 100 protein binding sites that each bind to one of 10 ligands, we assessed CPDA's ability to differentiate between LBSs across protein groups. We graphically evaluated the effectiveness of the CDPA method using Multidimensional scaling (MDS) plots for all pairs of ligand groups in the Kahraman dataset. In the next step, we decomposed squared Euclidean distances between pairs of LBSs into diagonal and off-diagonal components to assess the role the variances (diagonals) and covariance structure (off-diagonals) play in quantifying differences between LBSs. We then evaluated the contributions of the diagonal and off-diagonal Mahalanobis distances to the total Mahalanobis distance between pairs of LBSs using linear regression. Our analysis revealed that MDS plots could visually distinguish between most group pairings, with only six exceptions out of 36 possible combi-

nations. As our main goal, we explored the contribution of diagonal and off-diagonal elements to the total squared Euclidean distances and Mahalanobis distances. Results of the decomposition of Euclidean distances showed significant variability in their contributions across different pairings, but all results suggest that both diagonal and off-diagonal elements provide significant information about the structural information of LBSs. The dominant component the diagonal and off-diagonal Mahalanobis distances also showed considerable variability, as each type of component was dominant in different circumstances. Furthermore, in most cases, there are significant contrasts in the results when changing the reference group in the same group of pairs. Considering all these factors our approach to quantifying the individual contributions of diagonal and off-diagonal Mahalanobis distances to the total Mahalanobis distances suggests that the usage of either solely diagonal or solely off-diagonal components as a simpler encoding way of information of protein binding sites is not feasible. While diagonal distances often exhibited a stronger contribution, off-diagonal distances also provided essential contributions in most cases. However, the uneven pattern in contributions and non-negligible contribution of both diagonal and off-diagonal elements to the total distances of covariance matrices in most group pairing suggest that simplifying the encoding of structural information to either diagonal or off-diagonal elements alone may not be adequate, and thus we should utilize the full covariance matrix encoding.

6 *Circular Regression*

Pengyuan Chen, University of Kentucky

Abstract: Directional data has received increasing attention across a large number of scientific fields. In particular, such data assume some notion of an underlying circular distribution, which is characterized by some form of angular or degree direction. Naturally, modeling with such distributions when observed covariates are present necessitate the use of regression methods. However, circular variables have some specific characteristics which are different from linear variables, so traditional linear models need an appropriate transformation to become circular models. This paper extends the simple circular-circular regression model and the circular-linear model into multiple circular-circular regression models, and models based on both circular and linear covariates. We further develop a degree-determination algorithm that is used in the aforementioned models. This algorithm makes use of classic dimension reduction methods (principal component analysis and partial least squares) applied to multivariate circular regression models. Performance of our methods are investigated and compared based on both simulated and real datasets.

7 *Scalable K Sample Test using Empirical Likelihood*

Jeremiah Tella, Georgia State University

Abstract: The goal of the paper is to find a fast and efficient way of testing the equality of multiple samples without having to make any assumptions about the distribution of the samples, especially when the size of each sample is large. For example, we may have data about the effectiveness of different vaccines and be interested in knowing if the vaccines are equally effective or not. In statistics, this is called a k-sample test. So, we combined a method called 'divide and conquer' with another method called empirical likelihood, and we developed a new test that can correctly detect if the samples are from the same distribution. Our test was correctly able to detect that the samples are from the same distribution more than 95% of the time and was correctly able to detect that they are not from the same distribution more than 80% of the time. Our test is comparable to other available tests in this regard. However, it took many days for other methods to produce results, whereas our method produced results in minutes.

8 ***A Group Testing Based Exploration of Age-varying Factors in Chlamydia Infections among Iowa Residents***

Yizeng Li, University of South Carolina

Abstract: Group testing, a method that screens subjects in pooled samples rather than individually, has been employed as a cost-effective strategy for chlamydia screening among Iowa residents. In efforts to deepen our understanding of chlamydia epidemiology in Iowa, several group testing regression models have been proposed. Different than previous approaches, we expand upon the varying coefficient model to capture potential age-varying associations with chlamydia infection risk. In general, our model operates within a Bayesian framework, allowing regression associations to vary with a covariate of key interest. We employ a stochastic search variable selection process for regularization in estimation. Additionally, our model can integrate random effects to consider potential geographical factors and estimate unknown assay accuracy probabilities. The performance of our model is assessed through comprehensive simulation studies. Upon application to the Iowa group testing dataset, we reveal a significant age-varying racial disparity in chlamydia infections. We believe this discovery has the potential to inform the enhancement of interventions and prevention strategies, leading to more effective chlamydia control and management, thereby promoting health equity across all populations.

9 ***BiGER: Bayesian Rank Aggregation in Genomics with Extended Ranking Schemes***

Kaiwen Wang, Southern Methodist University

Abstract: With the rise of large-scale genomics and proteomics studies, large gene lists targeting specific diseases are increasingly common. While evaluating each study individually gives valuable insight on the specific sample and study design, the wealth of available evidence in the literature calls for robust and efficient meta-analytic methods.

Crucially, the diverse assumptions underlying different studies, such as the uncertain nature of top-unranked and unreported genes, require a flexible but rigorous method for aggregation. To address these issues, we propose BiGER, a fast and accurate Bayesian rank aggregation method for the inference of true latent rankings. Using a Bayeasian hierarchical framework and variational inference, BiGER efficiently aggregates large-scale gene lists with good accuracy, while the variance estimation procedure provides valuable insights on source reliability for researchers. Using both simulated and real datasets, we show that BiGER is a useful tool for reliable meta-analysis.

10 *SCIntRuler: Guiding the Integration of Multiple Single-cell RNA-seq Datasets with a Novel Statistical Metric*

Yue Lyu, University of Texas Health Science Center at Houston

Abstract: The growing number of single-cell RNA-seq (scRNA-seq) studies highlights the potential benefits of integrating multiple datasets, such as augmenting sample sizes and enhancing analytical robustness. Inherent diversity and batch discrepancies within samples or across studies continue to pose significant challenges for computational analyses. Questions persist in practice, lacking definitive answers: Should we use a specific integration method or opt for simply merging the datasets during joint analysis? Among all the existing data integration methods, which one is more suitable in specific scenarios? To fill the gap, we introduce SCIntRuler, a novel statistical metric for guiding the integration of multiple scRNA-seq datasets. SCIntRuler helps researchers make informed decisions regarding the necessity of data integration and the selection of an appropriate integration method. Our simulations and real data applications demonstrate that SCIntRuler streamlines decision-making processes and facilitates the analysis of diverse scRNA-seq datasets under varying contexts, thereby alleviating the complexities associated with the integration of heterogeneous scRNA-seq datasets.

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

Xiaoxi Zhang, University of Florida

Abstract: 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.

12 *Diffusion Process in Civil Infrastructure Network*

S M Mustaquim, The University of Texas at El Paso

Abstract: In a civil infrastructure system, failure in one component can cause failures in other components resulting in a cascading failure. We evaluate the transmission of the failures among components by considering the infrastructure system as a complex network, where nodes represent the components and edges represent the relationships among the components. Whether and how things diffuse from one network component to another depends on several factors, e.g., timing in diffusion, and interaction among different nodes. In this study, we conduct an extensive simulation study and investigate how the network structure, in particular, the local geometry of the network, affects the transmission or diffusion process. In a case study, we evaluate the resilience of real-world power grids under cascading failure generated by diffusion processes.

13 *Empirical Likelihood Inference for Mean Time to Failure Order*

Maxime Bouadoumou, Georgia state university

Abstract: In this paper, an empirical likelihood inference of the mean time to failure order function is proposed. Confidence intervals are compared through simulation studies in terms of coverage probability. We illustrate the proposed procedure using two real data sets.

14 *Novel Empirical Likelihood Method for the Cumulative Hazard Ratio under Stratified Cox Models*

Dazhi Zhao, Georgia state university

Abstract: In clinical studies, how to evaluate the treatment effect is a crucial topic. Nowadays, the ratio of cumulative hazards is often applied to accomplish this task, especially when those hazards may be nonproportional. The stratified Cox proportional hazards model, as an important extension of the classical Cox model, has the ability to flexibly handle nonproportional hazards. In this paper, we propose a novel empirical likelihood (EL) method to construct the confidence interval for cumulative hazard ratio under stratified Cox model. The large sample properties of the proposed profile EL ratio statistic are investigated, and the finite sample properties of the EL-based estimators under some different situations are explored in simulation studies. The proposed method was finally applied to perform statistical analysis on a real world dataset on the survival experience of patients with heart failure.

15 *Image Processing with Optimally Designed Parabolic Partial Differential Equation*

Qiuyi Wu, University of Rochester

Abstract: In imaging denoising tasks, brain imaging data like functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) scans often contain noise and artifacts. Kernel smoothing techniques are essential for smoothing these images and play a pivotal role in brain imaging analysis. While kernel smoothing has been extensively studied in statistics, certain challenges remain, especially in the multi-dimensional landscape. Many existing methods lack adaptive smoothing capabilities and numerical flexibility in high dimensional setting, hindering the achievement of optimal results. To address this, we present an efficient adaptive General Kernel Smoothing-Finite Element Method (GKS-FEM). This method exploits the equivalence between GKS and the general second-order parabolic partial differential equation (PDE) in high dimensions. Utilizing the Finite Element Method (FEM), we discretize the PDE, leading to efficient and robust numerical smoothing approaches. This study establishes a bridge between statistics and mathematics. The new method is designed to ensure efficiency and stability in high-dimensional scenarios. The primary applications for this work are in image processing and denoising tasks, which play an essential role in neuroimaging studies.

16 *Harmonizing Healthcare: The Art and Statistics of Consensus Building*

Joshua Cook, University of West Florida

Abstract: In the complex realm of medicine, it's essential to acknowledge that even seasoned practitioners may require guidance, especially when faced with uncertain or risky medical scenarios where traditional research methodology is deemed not feasible or unethical. Consensus statements are crucial tools for synthesizing expert opinions in such situations, offering a collective direction where singular expertise might be insufficient. However, there are several types of consensus studies, each with their own design schema and statistical guidelines that require deep understanding to ensure the validity and reliability of study outcomes. Our study provides a succinct guide to the intricacies involved in crafting these statements, addressing the definition of consensus, the optimal number of expert participants, and the balance between agreement and discussion rounds. We highlight the role of iterative feedback and the challenge of expert retention, supported by R simulations that assess the parameters that influence consensus achievement. Our conclusions serve as practical advisories for project managers and protocol writers, emphasizing that the process of reaching a consensus is not only iterative and collaborative but also integral to advancing medical practice and knowledge.

Abstracts for NSF/Harshbarger Student Posters

17 *Characterizing Interim Futility Analyses in Phase III Clinical Trials: A Systematic Review of Current Practices*

Corinne McGill, Medical University of South Carolina

Abstract: The longitudinal nature of large, phase III clinical trials often necessitates interim analysis of the primary efficacy outcome prior to trial conclusion. Interim futility analyses aim to prevent wasted resources and to avoid participant exposure to potentially ineffective treatments in the absence of a treatment effect. The pre-specification of futility stopping rules allows for objectivity in determining whether a trial's data and safety monitoring board should consider stopping the trial prior to its planned conclusion. However, specifically for trials with time-to-event outcomes and variable follow-up, consensus in the literature on best practices for selecting stopping rules is lacking. In an effort to better understand current practices in the design of interim futility analyses, a systematic review of the clinical trial literature, subset by primary outcome type (i.e., continuous, binary, time-to-event), was conducted. The review captured: common statistical methods used for futility stopping (e.g., conditional power, predictive power, error spending), specific thresholds utilized for each method and their frequency, and the proportion of trials that stopped early for futility under each threshold. PubMed, Scopus, Cochrane Central Register of Controlled Trials, and CINAHL were searched for candidate studies/manuscripts using the following terms: "randomized" or "randomised" with "futility", "interim", "phase 3" or "phase III", and not "noninferiority" or "non-inferiority". Results were filtered for English language, journal articles, not conference proceedings, and publication date between January 2019 and April 2024. Duplicates were then removed, and manuscripts were screened (N=123) before assessing eligibility (N=99). Manuscripts were deemed eligible if they described a phase III, superiority, prospective randomized controlled trial with at least one planned interim analysis for futility. 69 unique studies met eligibility criteria and were included in this review. The majority (65.2%) of these studies employed time-to-event primary endpoints, with a notable 75.6% of these utilizing event-based timing for interim analyses. The most prevalent method for determining futility stopping was through error spending functions, identified with a relative frequency of 30.4%. Among these functions, O'Brien-Fleming/O'Brien-Fleming-like approaches were the most frequently adopted. Conditional power followed with a relative frequency of 18.8%, with commonly utilized thresholds of 10% and 20%. While this review provides insight into the current practices in interim futility analyses, the imperative to establish definitive best practices remains unmet.

18 *Federated learning with GLMs under Parameter Heterogeneity*

Bhaskar Ray, North Carolina State University

Abstract: Federated Learning (FL) is an emerging framework for collaborative learning where a set of clients communicate via a central server to train statistically accurate models under privacy constraints. Under heterogeneity in the clients' data-generating distributions, several studies have shown that a common model learned globally may not generalize well with respect to the local distributions of the clients. In this study, we revisit the challenge posed by data heterogeneity in the context of FL for parameter estimation in a broad class of Generalized Linear Models (GLMs). We propose a novel statistical decision-theoretic approach that helps each client determine the subset of clients that it should collaborate with to balance the trade-off between the reduction in variance afforded by federation, and the bias introduced by data-heterogeneity. To achieve this, we construct statistical tests of similarity at the server that incorporate finite-sample properties of Maximum Likelihood Estimators for GLMs in tandem with tail bounds for stochastic optimization algorithms. Our overall approach enables each individual client to provably benefit from collaboration, as quantified by an improvement in its relative error-bounds that scales with the number of similar clients. Furthermore, such benefits do not come at the expense of additional bias terms due to heterogeneity. Our theoretical results are additionally validated using simulation studies.

19 *Jackknife Empirical Likelihood for Chatterjee Rank Correlation Coefficient*

Tope Amusa, Georgia State University

Abstract: In this paper, we propose using Jackknife Empirical Likelihood (JEL) and its variation for Chatterjee's rank Correlation Coefficient (CCC). CCC is a rank correlation coefficient that estimates a population quantity while remaining asymptotically normal under independence. However, it has been proven to be rate suboptimal compared to other recognized rank correlations. CCC has been demonstrated to adapt to the data's complex structure, making it a potentially useful tool. This study applies the Jackknife empirical likelihood (JEL) to estimate confidence intervals for the Chatterjee rank Correlation Coefficient and compares coverage probability and interval length for JEL and Normal Approximation (NA), Bootstrap, and Adjusted Jackknife Empirical Likelihood methods. Simulation studies are carried out to assess the suggested estimators' performance. Data are simulated for Correlation values between -0.75, -0.5, 0, 0.5, to 0.75 from normal and non-normal distributions of varying skewness. Simulation results showed that JEL methods perform better than the NA methods. Also, Bootstrap showed inconsistent coverage probability. The JEL methods have the widest confidence intervals in most cases.

20 *Predicting Dengue Incidence In The Dominican Republic Using Climate Data*

Sahil Chindal, Virginia Commonwealth University

Abstract: Dengue is a mosquito-borne disease prominent in tropical and subtropical regions of the world and has been emerging in temperate areas. Dengue is endemic to the Dominican Republic, where outbreaks have been occurring for the last four decades.

Most provinces in the Dominican Republic have a tropical climate with abundant rainfall during rainy seasons, providing ample resources for mosquito proliferation and growth, which creates an environment favorable for dengue transmission. Using climate data and dengue case data, we aim to determine which climate factors are associated with dengue cases. Additionally, due to the potential lags between climate and dengue trends, we seek to characterize the lags between variables associated with climate and cases. Using machine learning methods, we analyze the temporal dynamics of dengue spread by estimating the parameters of the SIR-type model framework. We validate our model using the historical data for dengue cases in the Dominican Republic. We will use our results as part of a predictive model to forecast the spread of dengue cases between various provinces of the Dominican Republic and integrate our model into an early warning system that predicts outbreaks and informs public health and mosquito control policies in the Dominican Republic.

21 *Addressing Duplicated Data in Point Process Models*

Lingling Chen, University of Houston

Abstract: Spatial point process models are widely applied to point pattern data from various fields in the social and environmental sciences. However, a serious hurdle in fitting point process models is the presence of duplicated points, wherein multiple observations share identical spatial coordinates. This often occurs because of decisions made in the geo-coding process. For example, assigning representative locations (e.g., aggregate-level centroids) to observations when data collectors lack exact location information. Researchers often use ad hoc solutions for duplicated data, necessitating alterations to the data before conducting statistical analysis with prevalent spatial point process models like the Log-Gaussian Cox Process (LGCP) and associated inference techniques such as the Minimum Contrast (MC) method. This study proposes a Modified Minimum Contrast (MMC) method that adapts the inference procedure to account for the effect of duplicates without needing to alter data. The proposed MMC method is applied to LGCP models, with simulation results demonstrating the gains of our method relative to existing approaches. The MMC approach is then used to infer the spatial clustering characteristics of conflict events in Afghanistan (2008-2009).

22 *Trends and Disparities in HPV Vaccination Among U.S. Adolescents from 2017 to 2021*

Victor Agboli, Georgia State University

Abstract: Despite the availability of Human papillomavirus (HPV) vaccines, HPV remains a prevalent sexually transmitted infection in the United States., primarily causing cervical and other cancers. We analyzed HPV vaccination trends among U.S. adolescents aged 13 – 17 years from 2017 to 2021, both before and during the COVID-19 pandemic, aiming to identify vaccination gaps and develop targeted interventions to reduce

HPV-related cancer incidence and improve public health. Utilizing National Immunization Survey-Teen data, the analysis reveals significant increases in HPV vaccine initiation and completion rates over the five years, with initiation rates rising by 11.4% (from 65.5% to 76.9%) and completion rates by 13.1% (from 48.6% to 61.7%). However, it's crucial to note that the U.S. fell short of the national Healthy People 2020 target of 80% HPV vaccination coverage. Significantly, disparities persist across demographics and geography. Vaccination rates remain lowest among males, non-Hispanic white adolescents, and in states situated in the southern region. Furthermore, logistic regression analysis reveals the substantial influence of various demographic and socioeconomic factors on HPV vaccination uptake, with males, non-Hispanic white teens, those with highly educated and married mothers, individuals from higher socioeconomic backgrounds, larger households, fewer doctor's visits, no provider recommendations, and no health insurance is less likely to initiate and complete the HPV vaccination series. This research emphasizes the need to enhance the HPV vaccination rate, focusing on addressing disparities among underserved groups. The findings provide invaluable insights for designing interventions that can lead to improved public health outcomes. While the COVID-19 pandemic may have contributed to this trend through increased healthcare access, public health campaigns, and preventive healthcare emphasis, further analysis is essential to gauge its impact on HPV vaccine initiation and completion rates. By striving to close these gaps, public health stakeholders can work towards achieving optimal HPV vaccination rates and reducing the burden of HPV-related cancers in the United States.

23 *Estimate Intensity–Duration–Frequency Curves with Duration-dependent Generalized Extreme Value Distribution*

Jiyun Huang, Clemson University

Abstract: Intensity-Duration-Frequency (IDF) curves, often estimated using the GEV distribution, provide critical insights for engineers and hydrologists in characterizing and managing extreme precipitation events. The rarity of such extreme rainfall occurrences, resulting limited observations or absent for certain durations, poses challenges for estimation across duration. To overcome data scarcity and facilitate the construction of continuous IDF curves across durations, the duration-dependent GEV (d-GEV) model has been proposed. In this study, we assess the performance of two commonly employed functional forms of the d-GEV model through Monte Carlo simulations. These simulations reveal inadequacies in the modeling assumptions, prompting exploration of alternative functional form such as polynomial and spline regression. Applying these approaches to a CanRCM4 dataset comprising 35 statistically independent ensembles, yields accurate evaluation of model performance. Our findings indicate that while the widely adopted scaling d-GEV models demonstrate improved daily estimation with reduced variability, particularly for larger return periods, they fail to fully capture the relationship between intensity and duration compared to the proposed models. With comparable variability to the pointwise GEV model, and superior alignment between intensity and duration changes, we therefore recommend employing the spline and polynomial regression approaches for constructing continuous IDF curves.

24 *Hybrid Genetic and Simulated Annealing Algorithm for High-dimensional Linear Models with Interaction Effects*

Leiyue Li, University of Kentucky

Abstract: High-dimensional linear regression with interactions has found its wide applications in areas such as social science, bioinformatics, and public health where the interaction effects between variables/features are of interest. We present hySAINT, an R package that efficiently implements both genetic algorithms and simulated annealing to conduct variable selection in high-dimensional linear regression models with two-way interaction effects under different hierarchical structures (strong, weak, or no heredity). Details of the underlying notion behind the algorithm is provided to demonstrate the usage of the R package. Simulations are used to illustrate the performance and merits of the R package, hySAINT.

25 *Advancements in Image Edge Detection for Computer Vision*

Jiacheng Xu, University of Kentucky

Abstract: Image edge detection is a critical area in the field of computer vision. My research initially focused on enhancing a non-parametric edge detection method developed in late 90s. Subsequently, my efforts shifted towards neural network-based methods for edge detection. The goal of this transition was to refine these techniques to reduce the reliance on human-annotated ground truth data required for training the models. This approach aims to streamline the training process and improve the adaptability and efficiency of edge detection algorithms in practical applications.

26 *Convolutional Non-Homogeneous Poisson Process and its Application to Wildfire Ignition Risk Quantification for Power Delivery Networks*

Guanzhou Wei, Georgia Institute of Technology

Abstract: To quantify wildfire ignition risks on power delivery networks, the current practice predominantly relies on the empirically calculated fire danger indices, which may not well capture the effects of dynamically changing environmental factors. This research proposes a spatio-temporal point process model, known as the Convolutional Non-homogeneous Poisson Process (cNHPP), and applies the model to quantify wildfire ignition risks for power delivery networks. The proposed model captures both the current (i.e., instantaneous) and cumulative (i.e., historical) effects of key environmental processes (i.e., covariates) on wildfire risks, as well as the spatio-temporal dependency

among different segments of the power delivery network. The computation and interpretation of the intensity function are thoroughly investigated. We apply the proposed approach to estimate wildfire ignition risks on major transmission lines in California, utilizing historical fire data, meteorological and vegetation data obtained from the National Oceanic and Atmospheric Administration and National Aeronautics and Space Administration. A comprehensive comparison study is performed to show the applicability and predictive capability of the proposed approach.

27 *Optimal Sensor Allocation for Emission Source Detection with Linear Atmospheric Dispersion Processes*

Xinchao Liu, Georgia Institute of Technology

Abstract: This research considers the optimal sensor allocation for estimating the emission rates of multiple sources in a two-dimensional spatial domain. Locations of potential emission sources are known (e.g., factory stacks), and the number of sources is much greater than the number of sensors that can be deployed, giving rise to the optimal sensor allocation problem. In particular, we consider linear dispersion forward models and the optimal sensor allocation is formulated as a bilevel optimization problem. The outer problem determines the optimal sensor locations by minimizing the overall Mean Squared Error of the estimated emission rates over various wind conditions, while the inner problem solves an inverse problem that estimates the emission rates. Two algorithms, including the repeated Sample Average Approximation and the Stochastic Gradient Descent based bilevel approximation, are investigated in solving the sensor allocation problem. Convergence analysis is performed to obtain the performance guarantee, and numerical examples are presented to illustrate the proposed approach.

28 *Efficient Design of Simultaneous Controlled Experiments in the Presence of Subject Covariates*

William Fisher, Clemson University

Abstract: Controlled experiments are a form of experiment that allows one to analyze a proposed treatment by comparing responses between a treatment group and control group. Researchers or organizations may run multiple, separate controlled experiments simultaneously due to time, budget, or other considerations. In this setting, subjects often participate in more than one experiment, and thus responses across experiments coming from the same subject are not independent. In addition, outside of the treatment, certain subject covariates may be explanatory factors in the subject's response to each experiment. In this talk, we consider the optimal experimental design problem of allocating subjects to treatment or control when subjects must participate in multiple controlled experiments simultaneously and subject covariate information is available. The goal of the allocation is to provide precise estimates of treatment effects for each experiment.

Deriving the precision matrix of the treatment effects in this setting and using D-optimality as our allocation criterion, we propose a greedy algorithm to solve the D-optimality problem. The greedy algorithm decomposes the D-optimality problem into a sequence of subproblems, where each subproblem can be solved either through integer programming or a semi-definite programming based randomized algorithm that originates from the MAXCUT problem. We showcase the performance of our greedy algorithm through a simulation study.

29 *Analyzing Breast Cancer using Topological Data Analysis*

Kelie Marline Momo Nizegha, Clemson University

Abstract: Breast cancer is a health disease common in women globally and comes as a result of abnormal growth of cells in the breast tissue. There are more than 2.3 million cases of breast cancer that occur each year and according to WHO, in most countries, breast cancer is the first or second leading cause of female cancer deaths. Due to this fact, there is need to enhance more techniques for better understanding in order to improve detection or identification of cancer. Topological data analysis (TDA) is an approach for analyzing data using algebraic topology. The poster will showcase how we apply TDA to analyze breast cancer dataset, to get insights information in the data.

30 *Predictor-Informed Bayesian Clustering*

Md Yasin Ali Parh, University of Louisville

Abstract: In this project we are interested in performing clustering of observations such that the cluster membership is influenced by a set of covariates. To that end, we employ the Bayesian nonparametric Common Atom Model (CAM), which is a nested clustering algorithm that utilizes a (fixed) group membership for each observation to encourage more similar clustering of members of the same group. CAM operates by assuming each group has its own vector of cluster probabilities, which are themselves clustered to allow similar clustering for some groups. We extend this approach by treating the CAM group membership as an unknown latent variable determined by the covariate variable. Consequently, observations with similar predictor values will be in the same latent group and are more likely to be clustered together than observations with disparate predictors. We propose a pyramid group model that flexibly partitions the predictor space into these latent group memberships. This model is similar to the Bayesian regression tree process except that it uses the same splitting rule for at all nodes of the same tree depth. Additionally, we introduce a block Gibbs sampler scheme for our model that uses the truncation approximation to the stick-breaking representations to perform posterior inference. Our methodology is demonstrated in simulation and real data examples.

Abstracts for NSF/Harshbarger Student Posters

31 *Physical-informed Storm Surge Estimation*

Katherine Kreuser, Clemson University

Abstract: Storm surge is the unusual rise in sea level caused by a storm's winds pushing water onshore. Due to high damage to roads, buildings, and lives, accurate estimation of storm surge high quantiles (e.g., r-year return level) and associated uncertainty are essential. Purely data-driven approaches to these tasks pose a challenge, as the number of hurricanes occurring near any single location is limited. A physically-driven approach, utilizing high-fidelity hydrodynamic computer simulations, provides an alternative by leveraging well-known physics to model how the surge level responds to hypothetical storms, where these simulated storms are parameterized by a few key characteristics. This approach requires the following tasks: 1) estimate the joint distribution of storm characteristics; 2) emulate the computer model input-output relationship via surrogate models; and 3) integrate out the input distribution to obtain the surge output distribution, then determine the synthetic data for high quantile (e.g. 1-in-100 year return level) estimation. A case study of this approach in South West Florida using the computer model ADCIRC will be presented to illustrate the proposed workflow.

32 *A Bayesian Hierarchical Approach for Evaluating Dichotomization of Continuous Variables to Predict Binary Outcomes*

Everette Keller, Medical University of South Carolina

Abstract: Dichotomization of continuous variables has been widely criticized in the statistical community, because of the potential for loss of information and power to test hypotheses. The choice of cut-point used for dichotomizing a continuous variable can lead to bias, particularly when the cut-point is chosen arbitrarily, for example choosing the sample mean or median, which may not be the same in other independent data. However, dichotomizing continuous variables to predict binary outcomes is prevalent in medicine offering straightforward interpretations and easy-to-implement decision rules. For example, specific systolic and diastolic blood pressure cut-points are commonly used to classify hypertension and determine treatment recommendations. Despite statistical concerns regarding information loss and reduced power, there may be scenarios where this is acceptable for the increased interpretability and ability to make clear cut decisions based on a cut-point of a continuous variable. In this work, we propose a Bayesian statistical approach which jointly (1) estimates the optimal cut-point for dichotomization of a continuous variable and (2) tests if information loss under this optimal dichotomization is greater relative to the continuous variable while incorporating the uncertainty for the location of the optimal cut-point. We evaluate our proposed approach in simulation and apply it to a case study examining age at first cannabis use and risk of a psychotic experience in adulthood.

Abstracts for NSF/Harshbarger Student Posters

33 *Machine Learning Algorithms with Effective Feature Selection Accurately Predict Unplanned Readmission after Ankle Surgery*

Zecheng Ling, Emory University

Abstract: This study employs machine learning (ML) algorithms to predict unplanned 30-day readmission after ankle surgery, by utilizing a retrospective dataset of patients who underwent ankle surgery from 2012 to 2019. The dataset was acquired from the National Surgical Quality Improvement Program database. The study variables include the demographics (e.g., age, gender, race), preoperative comorbidities (e.g. diabetes, hypertension), preoperative vitals and laboratory measurements (e.g., body mass index, creatinine, white blood cell count, hematocrit) of ankle surgery patient population. The primary outcome was unplanned 30-day readmission. The project evaluated 33 variables using eXtreme gradient Boosting (XGBoost), Recursive Feature Elimination (RFE) with cross-validation (CV) ML techniques to determine the optimal subset of covariates that could help predict the target outcome more accurately. Five ML models (based on XgBoost, Random Forest, Support Vector Machine, Neural Networks and Logistic Regression methods) were built on the training data using 10-fold CV and validated on the test data. The relative importance of the risk factors, precision, recall, area under the receiver operating characteristics (AUC), calibration and Brier score were used to assess and quantify the performance of ML models. These models proved effective in identifying the critical risk factors such as, 'Hypertension Requiring Medication' and 'Chronic Obstructive Pulmonary Disease'. The findings highlight the significant potential of machine learning to improve the accuracy of predicting health outcomes, thereby enhancing clinical decision-making processes and optimizing patient management in the context of ankle surgery. This approach could lead to better resource allocation and reduced healthcare costs associated with postoperative complications and readmissions following ankle surgery.

34 *Unraveling the Impact of Fuzzy Similarity Algorithms on Missing Data Imputation of Heart Bypass Surgery Cohort*

Hong-Jui Shen, Emory University

Abstract: The prevalence of missing data in medical research significantly impacts the accuracy of health outcomes analysis, particularly in critical areas such as Heart Bypass Surgery. This study introduces a novel imputation method based on Fuzzy C-Means and Random Forest (FCRF) similarity learning technique, aimed at addressing these challenges by enhancing the imputation accuracy across various missing data mechanisms—Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The study utilizes retrospective electronic health records (EHR) data of patients who underwent isolated coronary artery bypass grafting (CABG) at Emory-affiliated hospitals from 2014 to 2019. The study cohort includes 59 variables

comprising of patient demographics (e.g. age, gender, race), preoperative comorbidities (e.g., diabetes, hypertension, pneumonia), pre-and postoperative vitals and laboratory values (e.g., body mass index, heart rate, blood pressure, creatinine, hemoglobin) that are directly correlated with the outcomes of CABG. The cohort's sample size is 8,552 and it includes different missing data categories (e.g., MAR, MCAR, and MNAR). Preliminary results indicate that while the FCRF method shows promise, particularly in complex MNAR situations, it can still be fine-tuned. The efforts to further improve the imputation accuracy and thus maximize the efficacy of FCRF are in progress. The conclusion of this research posits the FCRF imputation method as a significant advancement in data imputation, potentially enhancing the integrity and reliability of health research data, with future studies aimed at refining the method further to ensure robust statistical inferences.

35 *Indicator-based Bayesian Variable Selection for Gaussian Process models in Computer Experiments*

Fan Zhang, Arizona State University

Abstract: Gaussian process (GP) models are commonly used in the analysis of computer experiments. Variable selection in GP models is of significant scientific interest but existing solutions remain unsatisfactory. For each variable in a GP model, there are two potential effects with different implications: one is on the mean function, and the other is on the covariance function. However, most of the existing research on variable selection for GP models has focused only on one of the effects. To tackle this problem, we propose an indicator-based Bayesian variable selection procedure to take into account the effects from both the mean and covariance functions. A variable is defined to be inactive if both effects are not significant, and an indicator is used to represent the variable being active or not. For active variables, the proposed method adopts different prior assumptions to capture the two effects. The performance of the proposed method is evaluated by both simulations and real applications in computer experiments.

36 *Analyzing ICU Admissions Post-Laparoscopic Cholecystectomy Using SQL and Machine Learning Techiques*

Pengfei Lou, Emory University

Abstract: Laparoscopic cholecystectomy, a prevalent surgical procedure to remove the gallbladder, generally results in minimal postoperative complications. Nevertheless, a subset of patients requires intensive care post-surgery, underscoring the necessity for predictive analytics to forecast ICU admissions. Utilizing a comprehensive dataset from the MOVER database, which includes records from 39,685 patients and 64,354 surgeries, this study conducted extensive data cleaning via MySQL and imputed missing values using the most frequent values with Python. The dataset was refined to 1,411 records across 24 columns. A Random Forest model enhanced by Recursive Feature

Elimination (RFECV) and Synthetic Minority Over-sampling Technique (SMOTE) was applied to identify critical predictors of ICU admissions. Additionally, REFCV with 5-fold cross-validation and hyperparameter tuning with 3-fold cross-validation were employed to optimize model performance. Comparative analysis with SVM, logistic regression, and CNN was performed to ascertain the most effective model. Feature selection identified seven key variables significantly influencing model accuracy, including Body Mass Index (BMI), patient age (AGE), type of care received within the hospital(PATIENT_CLASS_GROUP), discharge disposition (DISCH_DISP_C), pre-surgery physical status (ASA_RATING_C), and liver-related lab test abnormalities (Abnormal Flag). Analysis revealed that middle-aged women, particularly those with either mild or severe systemic disease who show liver damage, and who have a BMI over 40, are more likely to require ICU care following surgery. Notably, medication usage did not significantly correlate with ICU admissions. The optimization of model hyperparameters and feature reduction elevated the ROC AUC score from 0.77 to 0.85, with Random Forest achieving the best performance. This study highlights the critical role of demographic, clinical, and laboratory factors in predicting ICU admissions following laparoscopic cholecystectomy. The insights gained not only enhance our understanding of the associated risk factors but also showcase the capacity of machine learning techniques to improve postoperative care strategies.

37 *Neural Net Group Testing*

Yu Huang, Clemson University

Abstract: Group testing is a methodology that involves combining individual specimens for disease detection, posing challenges in modeling due to unobserved true individual statuses and potential misclassifications in testing responses. While previous regression methods focused on initial pool responses or made restrictive assumptions about assay accuracy, this paper presents a novel approach using Neural Networks (NN) for modeling group testing data. Our method offers a robust alternative for estimating covariate effects and inferring assay accuracy probabilities simultaneously, thereby providing a comprehensive analysis. We demonstrate the effectiveness of our approach using real-world data and provide user-friendly R code for practical implementation.

Boyd Harshbarger, PhD

(Feb. 15, 1906 – May 20, 1998)

Boyd Harshbarger was one of the early pioneers of statistics in the United States. He received his bachelor's degree from Bridgewater College, and his M.S. in Mathematics from the University of Illinois, and an M.S. from Virginia Tech. His career at Virginia Polytechnic Institute began in 1931, where he founded one of the earliest (third oldest) Departments of Statistics in the country. In 1935, he organized and taught the first courses in statistics in the Department of Mathematics. He received a Rockefeller Fellowship and left to pursue his Ph.D. in 1940. He wrote his doctoral dissertation under the direction of distinguished Professor W.G. Cochran at Iowa State College and George Washington University, and returned to VPI in 1942.

The success of the VPI Department of Statistics was due to the foundation laid by Boyd Harshbarger. Boyd Harshbarger was the founder of the Statistical Laboratory at VPI in 1947 and the Department of Statistics (1949). He was the head of the Department of Statistics from 1949 to 1972, and was named professor emeritus in 1976.

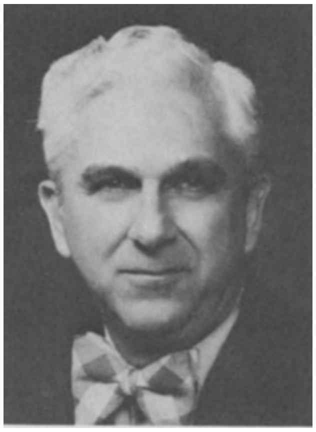
He received many honors in recognition of his services to the statistics profession and to the sciences in general. In addition to receiving an honorary doctorate from his alma mater, he was named a fellow of the American Association for the Advancement of Science (AAAS), the American Statistical Association (ASA), the Institute of Mathematical Statistics (IMS), and the Virginia Academy of Science. He was a charter member of the Biometric Society and president of the Eastern North American Region (ENAR) from 1956–1958. The VAS honored his work with the J. Shelton Horsley Research Award (1946), with the Ivey F. Lewis Distinguished Service Award (1966), and with honorary life membership. Harshbarger organized the Statistics Section of the VAS, and served as its secretary (1945-1946), vice-chair (1947) and chair (1948).

With respect to today's Southern Regional Council on Statistics, a Committee on Statistics was coordinated through the Southern Regional Education Board, and consisted of representatives from 14 southern states, from Maryland to Texas. (This was the precursor to today's SRCOS.) Their charge was to promote statistics graduate programs in the southern colleges and universities. Boyd Harshbarger was the first chairman of this committee (1961–1962). In 1954, Harshbarger organized and supervised the first Regional Statistical Summer Season (the first "Summer Research Conference") at VPI, bringing together scientists from five continents and students from 35 states of the US. Harshbarger influenced many students and colleagues with his enthusiasm and vision of statistics. The Boyd Harshbarger travel awards were giving in 1994 at the SRC in

Williamsburg, Virginia, supported entirely by the SRCOS. Beginning with the 2000 SRC, also in Williamsburg, support in the form of grants from federal agencies was obtained for this student award, along with the R. L. Anderson Award.

Richard L. Anderson, PhD

(April 20, 1915 – Jan. 19, 2003)



Richard "Dick" L. Anderson lived through the farm depression of the 1920s and the general depression of the 1930s. He missed two elementary grades, so that he was two years younger than his high school class mates. Income from clover seed bags was used to pay for his room and board at Depauw University, where he had a tuition scholarship, and he eventually went on to earn the AB degree in 1936, and his Ph.D. in Mathematics, Statistics, and Economics at Iowa State College in 1941. He began his academic career at North Carolina State College and rose from Instructor to full professor in his 25 years of service there to 1966. He was away from Raleigh in 1944–45 when he consulted with the Army and Navy at Princeton, in 1950–51 when he visited Purdue, and during 1958 when he spent time at the London School of Economics. In 1967, after a year as a visiting research professor at the University of Georgia, he took up the chairmanship of the newly established Department of Statistics at the University of Kentucky.

Professor Anderson had long-standing interests in experimental design, regression methods, variance components and time series analysis and their application to agricultural, industrial and operational problems. He took an active role in statistics in the US and internationally. He was a fellow of the American Statistical Association (ASA), the Institute of Mathematical Statistics, and of the American Association for the Advancement of Science. He was President of the Eastern North American Region of the Biometric Society, was a member of the International Statistical Institute, and was President of the ASA. He was a joint author of the well-known book Statistical Theory in Research.

Dick had a long association with the American Statistical Association, and was elected as a fellow in 1951, president in 1982, and received a Founder's Award in 1992. He was also a fellow of the American Association for the Advancement of Science and the Institute for Mathematical Statistics. He was a member of the International Statistical Institute, and the International Biometric Society.

Dick Anderson was one of the panel of original organizers of the Regional Statistical Summer Season (the precursor to the Summer Research Conference) and was a long-time member of SRCOS, and served as president of SREB-COS from 1975–1977. In recognition of his many contributions to the advancement of statistics in the South and nationally, graduate student travel awards were made annually in honor of Richard L. Anderson. The first awards were giving in 1994 at the SRC in Williamsburg, Virginia, supported entirely by the SRCOS. Beginning with the 2000 SRC, support in the form of grants from federal agencies was obtained for this student award, along with the Boyd Harshbarger Travel Awards.

M. Clinton Miller, PhD

(Aug. 28, 1932 – Feb. 10, 2004)

Dr. Miller received his high school diploma at Oklahoma Military Academy, where he also began his college work. He graduated from the University of Oklahoma with a degree in chemistry, and completed the first year of medical school at the Oklahoma University School of Medicine. His medical career goal was temporarily interrupted by a diving accident that left him a quadriplegic at the age of 21. Combating difficult odds, he soon returned to the University of Oklahoma, where he earned a master's degree in mathematics and to the Oklahoma Medical School, where he received a doctorate in biostatistics. He then did a number of post-doctoral fellowships and seminars.

Dr. Miller's long and distinguished career included teaching and research at the Medical Schools of Oklahoma, Tulane University in New Orleans, and the Medical University of South Carolina, where he established the Department of Biometry. During his 26 year tenure, he authored or co-authored 19 books and wrote or contributed to more than 81 scientific articles and publications. He worked on committees and projects as a volunteer and a consultant for countless health agencies, institutes, universities, clinics, and pharmaceutical companies, Departments of Vocational Rehabilitation and the Food and Drug Administration. He was also a fellow of the American Statistical Association.

Clint Miller was one of the panel of original organizers of the Regional Statistical Summer Season (the precursor to the Summer Research Conference) and was a long-time member of SRCOS, and served as president of SRCOS from 1981-1983. In his honor, the first annual Clint Miller Award for Outstanding Student Presentation was given at the 2004 Summer Research Conference in Blacksburg, Virginia, to Amy Bardeen and Thomas McCoy of Clemson University for their poster presentation of “Effects of Endometrial Growth Patterns and Embryo Transfer Time of Pregnancy Rates in an Assisted Reproduction Technology Program.”

Michael H. Kutner, PhD

Dr. Kutner received his master's degree in statistics from Virginia Polytechnic Institute and State University in 1962. He taught mathematics and statistics at the College of William and Mary from 1962 through 1967. In 1967, Dr. Kutner received a National Science Foundation Scholarship to pursue his doctoral training at Texas A&M University. After obtaining his doctoral degree in 1971, he took an Assistant Professorship tenure-track position in the Department of Statistics and Biometry at Emory University. Dr. Kutner was appointed Chair of the Department of Biostatistics when Emory formed the School of Public Health in 1990. He was appointed as inaugural Associate Dean for Academic Affairs in 1990 as well. In 1994, Dr. Kutner was recruited by the Cleveland Clinic Foundation as the Chair of the Department of Biostatistics and Epidemiology. Dr. Kutner returned to Emory University in 2000 as Professor of Biostatistics. In 2002, Dr. Kutner became Interim Chair of the Department of Biostatistics in the Emory Rollins School of Public Health. In 2004, he was named Rollins Professor and Chair of the Department of Biostatistics.

He has co-authored two widely-used textbooks and has published over 170 peer-reviewed manuscripts primarily in top-tiered medical journals. His statistical research interests have been notably in the areas of linear models and clinical trials. Dr. Kutner has received numerous awards and honors. He received the Thomas F. Sellers, Jr. Award in 2008 for serving as an excellent role model and mentor from the School of Public Health. In 2011, he received the Dr. Charles R. Hatcher, Jr. Award from the Health Sciences Center for his lifetime of work exemplifying excellence in public health. Dr. Kutner was named a Fellow of the American Statistical Association in 1984, received the ASA Founders Award in 1996 and the W. J. Dixon Award for Excellence in Statistical Consulting in 2011. He received the Mu Sigma Rho for lifetime devotion to statistical teaching award in 2009. In 2002, Dr. Kutner received the Paul Minton Award for service to the profession from the Southern Regional Council on Statistics (SRCOS) and he served as SRCOS President from 2008–2009. Texas A&M's Department of Statistics honored Dr. Kutner in 1984 when he was awarded the H. O. Hartley Award for distinguished service to the profession. In 1997, Texas A&M inducted Dr. Kutner as one of its inaugural recipients into the "Academy of Distinguished Graduates." In 2013, SRCOS named the Junior Faculty Poster Session at the Summer Research Conference in his honor. Junior and isolated faculty have been eligible to apply for Kutner travel awards if they present a poster.

Paul Dixon Minton, PhD

(Aug. 4, 1918 – July 10, 2007)

Paul Minton was born in Dallas, TX and grew up in a family with three brothers during the Great Depression. He attended Southern Methodist University (SMU) on an "emergency scholarship" made available to promising candidates from Dallas who otherwise could not afford college. He earned his bachelor's and master's degrees from SMU, Department of Mathematics, with interruption to serve as a cryptanalyst for the FBI during World War II.

Minton continued his studies at North Carolina State University, earning his PhD in 1957 under the direction of Gertrude Cox. While there, he was exposed to R. A. Fisher, H. Hoetelling, W. Cochran, J. Wolfowitz, H. E. Robbins, R. L. Anderson, R. C. Bose, and many other well known statisticians.

Dr. Minton then returned to SMU Mathematics and built a program of statistics courses for students from a wide range of majors. These evolved into the formation of the Department of Statistics in 1961, with Dr. Minton as the founding chair. The PhD degree was instituted in 1966. During this time, Dr. Minton was also named the director of the first computer center at SMU, which housed the Univac 1103, one of the few large scientific computers then available. After chairing the new department for 10 years, Dr. Minton joined Virginia Commonwealth University in 1972 as Dean of the School of Humanities and Social Sciences, a position he held until 1978. In 1979, he formed the Institute of Statistics at VCU and served as its Director until 1987, before retiring in 1988.

Dr. Minton was an active leader in the profession and received numerous awards. He was named a Fellow of the American Statistical Association in 1968 and received the prestigious ASA Founders Award in 1991. He founded the North Texas Chapter of the American Statistical Association (ASA) and was also very active in the Virginia Chapter. Dr. Minton represented SMU on the Southern Regional Education Board Committee on Statistics (precursor of SRCOS) from 1963 to 1972, and chaired this committee 1968–1969. He also represented VCU on this committee from 1979 to 1987.

Dr. Minton was known for his dedication to service in all his roles. In the words of J. M. Davenport, Minton "...was a great champion of those who were in need of assistance, opportunity, and needed encouragement. He had this innate ability to meet you where you were, offer words of wisdom, give you the resources you needed, and push you in the right direction. He was truly a great teacher." In further recognition of his contributions, the SRCOS established the Paul D. Minton Service Award in 1992 to recognize outstanding service to the statistics profession.

References for all biographies

- "Dr. Clint Miller, former Biometry chair, dies," *The Catalyst Online*, Medical University of South Carolina Office of Public Relations, Feb., 2004.
- *The Spectrum*, Virginia Polytechnic Institute, **20**(32), June 4, 1998.
- J. C. Arnold, K. Hinklemann, G. G. Vining, and E. P. Smith, "Virginia Tech Department of Statistics," in *Strength in Numbers: The Rising of Academic Statistics Departments in the U.S.*, A. Agresti and X-L. Meng (eds.), Springer-Verlag, 2013.
- Anderson, R.L. (1982). "My Experience as a Statistician: From the Farm to the University," in *The Making of Statisticians*, J. Gani (ed.), Springer, 129-148.
- Taylor, R.L. and Padgett, W.J. (2006). "The Summer Research Conferences and the SRCOS: A Historical Perspective," *J. Statistical Computation and Simulation*, **76**(5):373-383, May, 2006.
- "Richard L. Anderson" in *Statisticians in History*, published on-line by the American Statistical Association, (ASA biography url). See `http://www.amstat.org/about/statisticiansinhistory`.
- Brock, D.B. (2008). "Building a Department," *Amstat News*, Sept., issue 375.
- Davenport, J.M. (2008). "Paul Dixon Minton: LSD for Statisticians from a Southern Gentleman", paper from Memorial Session, Joint Statistical Meetings.

We extend our warmest thanks to all participants and to the following organizers:

Officers

Dr. Madhuri Mulekar, SRCOS President, University of South Alabama
Dr. Kathrine Thompson, SRCOS President-Elect, University of Kentucky
Dr. Edward Boone, SRCOS Past-President, Virginia Commonwealth University
Dr. John Wierman, SRCOS Treasurer, Johns Hopkins University
Dr. Norou Diawara, SRCOS Secretary, Old Dominion University

Scientific Program

Dr. Whitney Huang, Clemson University
Dr. David Hitchcock, University of South Carolina
Dr. Qiong Zhang, Clemson University
Dr. Xinyi Li, Clemson University
Dr. Edsel Peña, University of South Carolina

ASA/Kutner Faculty Posters and NSF/Harshbarger Student Posters

Dr. Edward Boone, Virginia Commonwealth University

Local Arrangements

Dr. Whitney Huang, Clemson University
Dr. Deborah Kunkel, Clemson University
Dr. Brook Russell, Clemson University
Dr. Yu-Bo Wang, Clemson University
Dr. Shyam Ranganathan, Clemson University

Website Administration

Kathrine Thompson, University of Kentucky
Whitney Huang, Clemson University

The Mission of SRCOS

The mission of SRCOS is to promote the improvement of postsecondary education in statistical science, assist in the development of high quality statistics instruction in elementary and high schools, and promulgate educational activities that improve the quality of statistical practices. SRCOS fulfills this mission by fostering and facilitating cooperation among institutions in its membership region concerned with statistics education.

Specific examples of SRCOS activities in fulfilling its mission are providing forums for communication on effective approaches to solving common problems with statistical training, sponsoring (joint with ASA) annual summer research conferences, and maintaining an electronic network for sharing statistical information among its members.

2024 Officers

President	Madhuri Mulekar, University of South Alabama
President-Elect	Kathrine Thompson, University of Kentucky
Past-President	Edward Boone, Virginia Commonweath University
Treasurer	John Wierman, Johns Hopkins University
Secretary	Norou Diawara. Old Dominion University
Web Administrator	Kathrine Thompson, University of Kentucky & Whitney Huang, Clemson University

The 2024 SRCOS SRC program booklet was produced by the School of Mathematical and Statistical Sciences at Clemson University.

Co-hosts of the SRCOS 59th Summer Research Conference

School of
MATHEMATICAL AND
STATISTICAL SCIENCES
Clemson® University

UNIVERSITY OF
South Carolina