Statistics and Data Science Seminar Series for Spring 2024

January 26

Speaker: Jackson Lautier
Affiliation: Department of Mathematical Sciences, Bentley University
Title: A New Framework to Estimate Return on Investment for Player Salaries in the National Basketball Association
Abstract: The evaluation of financial investment decisions always begins with a measurement of realized returns. Difficulties arise when returns are nonmonetary, however. One such example is offering a monetary salary in exchange for playing basketball, such as in the National Basketball Association. Despite many proposals to measure the on-court performance of basketball players, there are very few studies that consider both salary and on-court performance simultaneously. We thus present a novel five-part framework to translate a basketball player's on-court performance into a series of cash flows for the purpose of estimating a contractual return on investment (ROI). Our framework relies on a novel performance-based per-game wealth redistribution measurement that is calibrated with logistic regression on player tracking data against team wins, a WinLogit. The cumulative nature of individual player statistics summing to team statistics directly leads to pleasing statistical properties, such as the sum of player game logits equating to the team's game logit. We also find a maximum likelihood estimate for the WinLogit. We further demonstrate that centered player statistics allow for player calculations to be directly normalized to an average or replacement player, often desirable in sports analysis. We find WinLogit is more accurate than a per-game version of Win Score and Game Score at estimating team total wins and relative team rankings. We present ROI calculations based on WinLogit, Win Score, and Game Score. All results are presented for the 2022-2023 NBA regular season, including by position. For the purposes of replication, all data and code have been made available on a public repository.

February 2

Speaker: Seong Kim
Affiliation: Department of Mathematical Data Sciences, Hanyang University
Title: Bayesian Inference and Applications for Zero-Inflated Distributions
Abstract: Analysis of discrete data is frequently conducted in diverse fields, including natural sciences, social sciences, public health, and other disciplines. The binomial and Poisson distributions are perhaps commonly used in utilizing discrete data. For instance, it would be interesting to check the number of defective products in total production; to observe how many earthquakes will occur in one year; or to see how many home runs can be produced by a baseball batter in each game. These count data possess a considerable number of excessive zeros, hindering analysis with the regular binomial and Poisson distributions. Under these circumstances with excessive zero patterns, zero-inflated models would be a remedy to circumvent loss of information or tendencies of biased estimators. Prior elicitation has been one of the meaningful issues in both objective and subjective Bayesian inferences in which the prior distribution could reflect uncertainties about parameters before data are observed. In this talk, several zero-inflated models associated with Poisson, binomial, and bivariate binomial distributions are analyzed. Both noninformative and informative priors in conjunction with each model are intensively presented. Several real datasets are analyzed to support theoretical results.

February 9

Speaker: Charles South
Affiliation: Department of Statistics and Data Science, Southern Methodist University
Title: A Basketball Paradox: Exploring NBA Team Defensive Efficiency in a Positionless Game
Abstract: In the last decade, the offensive and defensive philosophies employed by teams in the National Basketball Association (NBA) have changed substantially. As a result, most players can no longer be classified into

only one of the five traditional positions (PG, SG, SF, PF, C) and instead spend a percentage of their playing time at multiple positions, making positional data compositional. Further, given the desirability for versatile players, an argument can be made that traditional positions themselves are archaic. Using data from the 2016-17, 2017-18, and 2018-19 seasons, I explore how Bayesian hierarchical models can be used to estimate team defensive strength in three ways. First, only considering players classified by their majority traditional position. Second, by using compositional traditional positional data. Third, using compositional data from modern positions (archetypes) defined by fuzzy k-means clustering. I find that the fuzzy k-means approach leads to a modest improvement in both the root mean squared error and median 95% posterior predictive interval width for the test data, and, more importantly, identifies 11 modern archetypes that, when combined, are correlated with team win total and adjusted team defensive rating. The modern archetype compositions can be used by stakeholders to better understand team defensive strength.

February 9

Speaker: Larry D'Agostino
Affiliation: Stellantis Financial and R User Group
Title: Knowledge and Insight From Career in Data Science
Abstract: A career in data science and analytics can lead to many places. This talk is one example of many of those experience. The goal of the talk is to discuss four main topics and advice on getting into a career in data science. The hope is to impart some wisdom on the future generation in this field of study. The four parts of the talk are 1) Starting Your Career, 2) Marketing Yourself, 3) Assembling Your Toolbox, and 4) Developing Data Skills.

February 23

Speaker: Alessandro Rinaldo
Affiliation: Department of Statistics and Data Sciences, University of Texas at Austin
Title: Sequential Change-point Detection for Network Data
Abstract: We study the change point detection settings in which we are presented with a stream of independent, labeled networks on a fixed node set, whose distributions are piece-wise constant over time. Our goal is to determine, as soon as we acquire a new observation, whether the data collected so far have provided sufficient evidence to infer that the underlying distribution has changed at the present time or in the near past. For sequences of Bernoulli networks, we formulate polynomial time CUSUM-based procedures and derive high-probability bounds on the corresponding detection delays with an explicit dependence on the network size, the entrywise and rank sparsity, and the magnitude of the change. We complement our analysis with minimax lower bounds, which we show are realized by NP-hard procedure. We also consider change point detection for sequences of multilayer random dot product networks with fixed and static latent positions but time-varying connectivity matrices. To handle this more complex and subtle scenario, we develop a sequential change point algorithm based on tensor methods and analyze its properties.

March 1

Speaker: Difeng Cai
Affiliation: Department of Mathematics, Southern Methodist University
Title: Data-Driven Kernel Matrix Computations: Geometric Analysis and Scalable Algorithms
Abstract: Dense kernel matrices arise in a broad range of disciplines, such as potential theory, molecular biology, statistical machine learning, etc. To reduce the computational cost, low-rank or hierarchical low-rank techniques are often used to construct an economical approximation to the original matrix. In this talk, we consider general m-by-n kernel matrices associated with possibly high dimensional data. We perform analysis to provide a straightforward geometric interpretation that answers a central question: what kind of subset is preferable for skeleton low-rank approximations. Based on the theoretical findings, we present scalable and robust algorithms for approximating

general kernel matrices that arise in astrophysics, kernel ridge regression, Gaussian processes, etc. The efficiency and robustness will be demonstrated through extensive experiments for various datasets, kernels and dimensions.

March 22

Speaker: Michael Gallaugher
Affiliation: Department of Statistical Science, Baylor University
Title: Cluster Weighted Models with Heavy Tailed Matrix Variate Distributions
Abstract: Cluster weighted models (CWMs) are powerful clustering devices used in many regression-type analyses. Unfortunately, real data often present atypical observations that make the commonly adopted normality assumption of the mixture components inadequate. Thus, to robustify the CWM approach in a matrix-variate framework, we introduce ten CWMs based on the matrix-variate t and contaminated normal distributions. Furthermore, once one of our models is estimated and the observations are assigned to the groups, different procedures can be used for the detection of the atypical points in the data. Parameter estimation via an ECM algorithm will be discussed, and the method will be applied to simulated data, as well as a real data analysis on greenhouse emissions data.

April 5

Speaker: Yulun Liu
Affiliation: Peter O'Donnell Jr. School of Public Health, UT Southwestern
Title: Network meta-analysis made simple: A composite likelihood approach
Abstract: Network meta-analysis, also known as mixed treatments comparison meta-analysis or multiple treatments meta-analysis, expands upon conventional pairwise meta-analysis by simultaneously synthesizing multiple interventions in a single integrated analysis. Despite the growing popularity of network meta-analysis within comparative effectiveness research, it comes with potential challenges. For example, within-study correlations among treatment comparisons are rarely reported in the published literature. Yet, these correlations are pivotal for valid statistical inferences. As demonstrated in earlier studies, ignoring these correlations can lead to inflated mean squared errors in estimates and inaccurate standard errors. In this talk, I will introduce a composite likelihood-based approach that guarantees accurate statistical inferences even without knowledge of these within-study correlations. The approach is computationally robust and efficient, with substantially reduced computational time compared to the state-of-the-science methods implemented in R packages. The proposed method has been evaluated through extensive simulations and applied to two important applications, including a network meta-analysis comparing interventions for primary open-angle glaucoma and another comparing treatments for chronic prostatitis and chronic pelvic pain syndrome.

April 26

Speaker: Pedro Maia
Affiliation: Department of Mathematics, University of Texas at Arlington
Title: Mathematical models and methods in computational neurology
Abstract: The emerging field of computational neurology provides an important window of opportunity for modeling of complex biophysical phenomena, for scientific computing, for understanding functionality disruption in neural networks, and for applying machine-learning methods for diagnosis and personalized medicine. In this talk, I will illustrate some of our latest results across different spatial scales spanning a broad array of mathematical techniques such as: (i) numerical methods for nonlinear PDEs for solving inhomogeneous active cable equations, (ii) spike-train metrics for quantifying information loss on compromised neural signals, (iii) applied dynamical systems for modeling biological neural networks, (iv) decision-making models, (v) applied inverse-problem techniques for finding the origins of neurodegeneration, and (vi) data methods in medical imaging.