IOWA | Department of Statistics and Actuarial Science

Hogg and Craig Lectures: The 50th Festival
April 28 – 29, 2023

Program

Friday, April 28
2520D University Capitol Centre

11:15 – 12:15 The Annual Hogg and Craig Cake
Poster Session featuring current graduate students

Moderator: Joyee Ghosh

1. Sam Dannels
   "Creating Disasters: Recession Forecasting with GAN-Generated Synthetic Time Series Data"
2. Shamriddha De
   "Robust Bayesian Variable Selection Using a Hyperbolic Error Model"
3. Zhenhan Fang
   "Density Estimation Using Nonlinear Independent Components Estimation (NICE)"
4. Chenyang Li
   "Maximum Likelihood Estimation of a First-Order Antedependent Model for Longitudinal Count Data"
5. Ian Lundy
   "A Comparison of Frequentist and Bayesian Model Selection Methods"
6. Anh Nguyen
   "Selecting Markov Chain Monte Carlo Algorithms with Kernel Stein Discrepancy"
7. Collin Nill
   "ZIEL: A Simple Yet Powerful Approach to Zero-Inflated Data"
8. Sreya Sarkar
   "Tree-Based Regression for International Classification of Diseases"
9. Nathan Tansey
   "Optimal Acoustics: Measuring the Robustness of Sounds to Distortion"
10. J. C. Thomas
    "Cancer Pathology Classification with Radiomic Imaging Covariates"
11. Yilin Wang
    "Online Statistical Inference and Conformal Prediction with Nonstationary Streaming Data"

12:15 – 12:20 Opening Remarks
Kung-Sik Chan, DEO of the Department of Statistics and Actuarial Science
Christopher Cheatum, Associate Dean of the College of Liberal Arts and Sciences

12:20 – 12:45 Awards Ceremony

Presented by Erning Li, Ambrose Lo, and Kung-Sik Chan

Student Academic Achievement Awards

Faculty Awards

Best Poster Awards

12:45 – 1:00 Break

1:00 – 2:00 **Hogg and Craig Lecture #1**

Dan Nettleton

"My Adventures in Sports Statistics, Beginning with Bob Hogg"

I was a student in an introductory mathematical statistics course Bob Hogg taught in the spring of 1992. Professor Hogg told many stories and seemed to have a lot of fun telling them. His enthusiasm for his stories and for statistics convinced me that I had made the right choice to pursue a graduate degree in statistics. One of Professor Hogg's examples used data from Major League Baseball to illustrate Simpson's Paradox. I was fascinated and interested in combining statistics with my enjoyment of sports. This talk will cover some of my work in the area of sports statistics, including my first refereed publication, which investigates homecourt advantage for the 1997 University of Iowa men's basketball team. Other topics in the presentation include (1) estimating win probability during competitions, such as National Football League games; (2) improving the Iowa High School Athletic Association's approach to ranking football teams; and (3) estimating victory probabilities for the canceled 2020 National Collegiate Athletic Association basketball tournaments.

2:00 – 2:15 Break

2:15 – 3:30 Panel Discussions on Hot Topics in Actuarial Science

Moderator: Barbara Hogg (Aon)

Panelists:

David Dillon (Lewis & Ellis)

Andy Ferris (Deloitte Consulting)

Marty Klein (Athene)

Larry Lickteig (Transamerica)

Dinner on your own

Saturday, April 29
Big Ten Theatre, 348 Iowa Memorial Union

8:30 – 8:45 Refreshments and assorted pastries

8:45 – 8:50 Opening Remarks
Kung-Sik Chan, DEO of the Department of Statistics and Actuarial Science

8:50 – 10:00 Departmental Faculty Research Showcase
Moderator: Dale Zimmerman
Speakers:
Lan Luo
"Online Statistical Inference with Dynamic and Dependent Streaming Data"
Streaming data refers to high-throughput recordings with large volumes of observations gathered sequentially and perpetually over time. Despite a large amount of work in the field of online learning, most of them are established under strong independent and identical data distribution, and very few target statistical inference. This talk will center around three key components in online statistical inference: (i) renewable updating, (ii) cross-batch dependence, and (iii) time-varying effects. I will first introduce how to conduct a renewable updating procedure, in the case of independent data batches, with a particular aim of achieving similar statistical properties to the offline oracle methods but enjoying great computational efficiency. Then I will discuss how we handle the dependence structure that spans across a sequence of data batches to maintain statistical efficiency in the process of renewable updating. Lastly, a dynamic weighting scheme will be integrated into the online inference framework to account for time-varying effects. This is joint work with Dr. Jingshen Wang from UC Berkeley and Dr. Emily Hector from North Carolina State University.

Sanvesh Srivastava
"Asynchronous and Distributed Data Augmentation for Massive Data Settings"
Data augmentation (DA) algorithms are widely used for Bayesian inference due to their simplicity. In massive data settings, however, DA algorithms are prohibitively slow because they pass through the full data in any iteration, imposing serious restrictions on their usage despite the advantages. Addressing this problem, we develop a framework for extending any DA that exploits asynchronous and distributed computing. The extended DA algorithm is indexed by a parameter $0 < r < 1$ and is called Asynchronous and Distributed (AD) DA with the original DA as its parent. Any ADDA starts by dividing the full data into k smaller disjoint subsets and storing them on k processes, which could be machines or processors. Every iteration of ADDA augments only an r-fraction of the k data subsets with some positive probability and leaves the remaining (1-r)-fraction of the augmented data unchanged. The parameter draws are obtained using the r-fraction of new and (1-r)-fraction of old augmented data. For many choices of k and r, the fractional updates of ADDA lead to a significant speed-up over the parent DA in massive data settings, and it reduces to the distributed version of its

parent DA when r=1. We show that the ADDA Markov chain has the desired stationary distribution under mild conditions on the parent DA algorithm. Coauthors: Jiayuan Zhou, Kshitij Khare (Department of Statistics, University of Florida)

Luke Tierney

"Adding a Pipe Operator to Base R"

This talk will briefly discuss the background and current state of the forward pipe operator recently added to the base R language.

Zhiwei Josh Tong

"The Vulnerability of Insurers to Interest Rate and Market Risks"

Life insurers and pension funds are exposed to long-term liabilities that are highly sensitive to changes in interest rates. To mitigate this risk, these institutions may implement a range of strategies, including holding a combination of short-term and long-term bonds, investing in the stock market, and taking fixed-leg positions in interest rate swaps. However, these strategies may be less effective and could potentially endanger insurance solvency during periods of high interest rates and under performance of equity market. We build and estimate a model for insurers' decision making and assess their vulnerability.

This is based on a collaboration with Zining Liu from CUFE and Chen Wan from ETH Zurich.

Boxiang Wang

"The Art of Transfer Learning: An Adaptive and Robust Pipeline"

Transfer learning is an essential tool for gaining information from auxiliary data resources to boost the performance of primary tasks. In this talk, I will introduce Adaptive Robust Transfer Learning (ART), a flexible pipeline of performing transfer learning for generic machine learning algorithms. The non-asymptotic learning theory of ART provides a provable theoretical guarantee to achieve adaptive transfer while preventing the so-called negative transfer. I will further introduce an ART-integrated-aggregating machine to output a single final model when multiple candidate algorithms are considered. The promising performance of ART will be demonstrated with extensive empirical studies for regression, classification, and sparse learning. A real-data analysis will be presented for the mortality study.

This is joint work with Chenglong Ye from University of Kentucky and Yunan Wu from University of Texas Dallas.

Xinyu Zhang

"Spectral Change Point Estimation for High Dimensional Time Series by Sparse Tensor Decomposition"

We study the problem of change point (CP) detection with high dimensional time series, within the framework of frequency domain. The overarching goal is to locate all change points and for each change point, delineate which series are activated by the change, over which set of frequencies. The working assumption is that only a few series are activated per change and frequency. We solve the problem by computing a CUSUM tensor based on spectra estimated from blocks of the observed time series. A frequency-

specific projection approach is applied to the CUSUM tensor for dimension reduction. The projection direction is estimated by a proposed sparse tensor decomposition algorithm. Finally, the projected CUSUM vectors across frequencies are aggregated by a sparsified wild binary segmentation for change point detection. We provide theoretical guarantees on the number of estimated change points and the convergence rate of their locations. We derive error bounds for the estimated projection direction for identifying the frequency-specific series that are activated in a change. We provide data-driven rules for the choice of parameters. We illustrate the efficacy of the proposed method by simulation and a stock returns application.

10:00 – 10:15 Break

10:15 – 11:15 **Hogg and Craig Lecture #2**

Dan Nettleton

"Who Is Winning? Determining Whether a Candidate Leads in a Ranked-Choice Election"

In an election with more than two candidates, it can be surprisingly complicated to determine whether a candidate is leading from the results of a survey. Even under the simplifying assumption of a simple random sample from an effectively infinite population, there are interesting statistical aspects to consider. Testing whether a particular candidate leads in the population involves a likelihood ratio test whose asymptotic null distribution is a chi-square mixture of the type arising in order-restricted inference. Complexity increases for elections in which each voter is asked to rank candidates on a ballot rather than simply choosing a most preferred candidate. In such ranked-choice elections, instant-runoff voting is often used to determine a winner. We explore how to test whether a particular candidate leads in the voting population based on candidate rankings provided by a sample of voters. We discuss a likelihood ratio test, an intersection-union test, and a simple Bayesian approach for evaluating whether a candidate leads a multi-candidate race in an election that uses ranked-choice and instant-runoff voting.

11:15 – 11:30 Group Photo – Please gather just outside the IMU North Entrance, which is near the Main Lounge, Room 180, in that area between IMU and Iowa Advanced Technology Laboratories (IATL).

Lunch on your own

1:30 – 3:15 Research Talks on Emerging Topics and Practice in Data Science

Moderator: Sanvesh Srivastava

Speakers:

Subhashish Chakravarty (Collins Aerospace)

"Emerging Topics and Practice in Data Science – An Aviation Perspective"

Data Science practice requires domain knowledge and a combination of skills acquired from a variety of academic disciplines. We will look at a few

of the current problems faced by the aviation industry and highlight elements of the cross-disciplinary approach towards addressing them.

Aaron Christ (United States Fish and Wildlife Service)

"Statistics Gone Wild: The 'Other' Biometrics"

Monitoring wildlife populations poses many challenges—even more with additional logistics and restrictions when dealing with remote environments. As with most messy data, solutions usually require a mixture of creativity and pragmatism in order to make timely management decisions. I will present a small selection of topics and issues encountered over my last two decades as a wildlife management agency biometrician. Ongoing projects and future directions will round out our adventure into Alaskan wildlife monitoring.

Kun Chen (University of Connecticut)

"Color Me Blue: A Journey Towards Data-Driven Suicide Prevention"

With the Zero Suicide initiative, preventing suicidal behavior among patients in the healthcare system has become a national priority. To tackle this issue, we have built an interdisciplinary research team consisting of behavioral scientists, public health experts, clinical doctors, computer scientists, and statisticians. By leveraging large-scale medical and socio-economic data from disparate sources, we aim to identify suicidal risk factors and predict suicidal behaviors to inform clinical decision-making and prevention strategies. In this presentation, we will briefly review our journey towards data-driven suicide prevention, with a focus on statistics and machine learning methodological development.

Dai Feng (Abbvie)

"Application of Artificial Intelligence and Machine Learning in Drug Development Life Cycle"

Ever since artificial intelligence (AI) was conceptualized in Alan Turing's work in the 1950s, the field of AI has progressed in various fields and more recently in drug development. By far, applications of AI in drug development have been largely focused on machine learning (ML). Comparing to the traditional statistical methods, AI/ML can tackle much larger data sets both structured and unstructured, identify hidden patterns, and make prediction of future outcomes. Once trained, they can parse data, continuously learn from them, and improve their performance. A recent landscape analysis based on drug and biologic regulatory submissions to the FDA from 2016 to 2021 shows that submissions with AI/ML components have increased rapidly in the past few years. In this presentation, I will present several use cases that demonstrate the power and growing acceptance of AI/ML in drug development life cycle. I will discuss challenges and future potentials as well.

Congrui Yi (Amazon)

"Personalized Long-Term Optimization with Progressive Learning from Day 1 to Day N"

For membership subscription services such as Amazon Prime, Statistics and Machine Learning (ML) based personalization is critical for driving

continuous growth in member acquisition, retention, and engagement. However, a unique challenge with this domain is the business need for optimizing meaningful, long-term outcomes that require several months to observe. Additionally, deep personalization can require millions of data points, and global expansion demands extensive development and maintenance efforts. All these factors incur substantial lead time and impede adoption of ML. Historically, solutions to this involve compromises of the desired outcome, such as optimizing for short-term proxies with multi-armed bandits, or deferring the use of a long-term modeling solution until a sufficient volume of matured outcome data has been collected. To address these problems, we proposed and developed a personalized recommender model that can progressively learn outcome patterns from Day 1 to Day N, starting with low data volume, and automatically scaling across countries. We demonstrated its effectiveness through simulations, offline and online experiments.

3:15 – 3:30 Break

3:30 – 5:00 Panel Discussions on Paths to Becoming a Successful Data Scientist

Moderator: Hai Liu (BioMarin)

Panelists:

Levent Bayman (Labcorp Drug Development)
Yeh-Fong Chen (FDA)
Rui Jin (Novartis)
Zhijiang (Van) Liu (Google)
Javier Porras (Zurich North America)
Bo Wang (Meta)

The North Room, 181 Iowa Memorial Union

6:00 Banquet Dinner – *Prior Reservation Required*

Closing Remarks

Kung-Sik Chan, DEO of the Department of Statistics and Actuarial Science
Barbara Hogg, Aon