Colorado/ Wyoming Chapter of the American Statistical Association
Spring Meeting – Friday April, 28th, 2023
Oracle, Broomfield Co.

Agenda

9:30-10:00 Donuts coffee
10:00-10:20 Frank Appiah, Oracle. "An overview of: A comparison of methods for predicting future cognitive status: mixture modeling, latent class analysis, and competitors"
10:20-10:40 Sven Serneels Gallop Data, Inc. Modeling Non-fungible tokens (NFTs)
10:40-11:00 Michael Creutzinger Statistics Department, Colorado State University "New Methods for Functional Outlier Detection"
11:00-11:20 Sandra Biller University of Wyoming. "Does your data mumble or sing? Tips on presenting data to non-statisticians."
11:20-11:40 Alex Hughes, Metro State University of Denver. "The Use of Artificial Control Groups in the Analysis of Retention among First-Year Freshmen"
11:40-13:00 Lunch
13:00-13:15 Chapter Business/ Elections/ David Young Awards
13:15-14:00 Steve Sain, Jupiter Intel. "Data science and assessing risk in a changing climate"
14:00-14:20 Cadet Allie Griffith & Cadet Michael Stanis , United States Air Force Academy. "Discover the potential of ChatGPTin a learning environment."
14:20-14:40 Erika Esquinca CU Anschutz Medical Campus Colorado School of Public Health METABOLOMICS PREDICTION FOR GENETIC DISCOVERY IN COPD USING METABOXCAN
14:40-15:00 Break
15:00-15:20 Troy Wixson Colorado State University Wildfire risk modeling with extremes
15:20-15:40 Mantautas Rimkus Colorado State University Classification methods for fault localization in future power grids
15:40-16:00 Danielle Dematteis CSU Statistics Department Distributed lag interaction model for modeling maternal exposure to air pollution
16:00-16:20 Connor Gibbs Colorado State University ECoHeN: A Hypothesis Testing Framework for Extracting Communities from Heterogeneous Networks

Abstracts

Frank Appiah, Oracle

An overview of: A comparison of methods for predicting future cognitive status: mixture modeling, latent class analysis, and competitors

This work compares various methods for using baseline cognitive performance data to predict eventual cognitive status of longitudinal study participants at the University of Kentucky's

Alzheimer's Disease Center. Cox proportional hazards models were used to examine time to cognitive transition as predicted by risk strata derived from normal mixture modeling, latent class analysis, and a one-standard-deviation thresholding approach. The results suggested three risk strata based on CERAD T scores: high, intermediate, and low risk. Cox modeling of time to cognitive decline based on posterior probabilities for risk stratum membership (from the normal mixture modeling method) yielded an estimated hazard ratio (HR) of 4.00 (1.53, 10.44) in comparing high risk membership to low risk; for intermediate risk membership versus low risk, the modeling yielded HR = 2.29 (0.98, 5.33). Other methods were also investigated. All methods for generating predictors of cognitive transition yielded statistically significant likelihood ratio statistics but modest concordance statistics. We concluded that the posterior probabilities from normal mixture modeling allow for risk stratification that is data-driven and modestly predictive of later cognitive decline. Incorporating other covariates may enhance predictions.

Sven Serneels, Gallop Data, Inc.

Non-fungible tokens (NFTs) constitute a novel financial asset class that has attained a market capitalization in excess of ten billion USD in just a few years. Therefore, it should not surprise that a host of financial services have been created for this segment, such as hedge funds that trade NFTs or lending services that take NFTs as a collateral. The success of each of these financial services depends on the algorithms they deploy and the quality of the data the latter are based upon. Owing to the nonfungible nature of these tokens, however, data related to NFTs pose unique modeling challenges not encountered in traditional financial markets. In this talk, a few such challenges will be discussed, such as low liquidity, high volatility and the presence of malicious market action, such as wash trading. A hint to mitigate these issues will be given.

Michael Creutzinger Statistics Department, Colorado State University

New Methods for Functional Outlier Detection. Functional data are data collected on a curve, or surface, over a continuum. The growing presence of high-resolution data has greatly increased the popularity of using and developing methods in functional data analysis (FDA). While functional data may have different characteristics from other data structures, the data analysis steps used with non-functional data are still relevant when applied to functional data (e.g. exploration, modeling, and inference). For example, even with functional data, there is a need to identify outliers prior to statistical analysis procedures. Existing functional data outlier detection methodology requires the use of a functional data depth measure, functional principal components, and/or an outlyingness measure like Stahel-Donoho. Although effective, these functional outlier detection methods may not be easily interpreted. I propose two new functional outlier detection methods. The first method, Practical Outlier Detection (POD), makes use of ordinary summary statistics (e.g. minimum, maximum, mean, variance, etc) to identify outliers. In the second method, I developed a Prediction Band Outlier Detection (PBOD) method that makes use of parametric, simultaneous, prediction bands that meet nominal

coverage levels. Both methods are compared to MS-Plot, Massive Unsupervised Outlier Detection, and Total Variation Depth outlier detection methods. In the preliminary results, POD performs as well, or better, than its counterparts in terms of specificity, sensitivity, accuracy, and precision. Similar results were found for PBOD, except for noticeably smaller values of specificity and accuracy than all other methods. I also present results using a data set studying world population growth since 1950.

Sandra Biller, University of Wyoming

Does your data mumble or sing? Tips on presenting data to non-statisticians While pursuing my MS in Statistics at the University of Wyoming, I have been working full time as an Associate Research Scientist with the Wyoming Survey & Analysis Center (WYSAC). I have observed that a critical component of statistical analyses is ensuring that results not only reach the right people but are also understood by those individuals. Effective data visualizations will engage an audience, facilitate understanding, and inform decision-making. Unfortunately, many chart defaults and common practices end up masking insights rather than enhancing them. In this presentation, I will: 1. Explain why effective charts matter; 2. Provide three best practices for data visualization; and 3. Illustrate how to easily transform a “bland” chart into one that is more focused and engaging.

Alex Hughes, Metro State University of Denver

The Use of Artificial Control Groups in the Analysis of Retention among First-Year Freshmen. An overview of: A comparison of methods for predicting future cognitive status: mixture modeling, latent class analysis, and competitors This work compares various methods for using baseline cognitive performance data to predict eventual cognitive status of longitudinal study participants at the University of Kentucky's Alzheimer's Disease Center. Cox proportional hazards models were used to examine time to cognitive transition as predicted by risk strata derived from normal mixture modeling, latent class analysis, and a one-standard-deviation thresholding approach. The results suggested three risk strata based on CERAD T scores: high, intermediate, and low risk. Cox modeling of time to cognitive decline based on posterior probabilities for risk stratum membership (from the normal mixture modeling method) yielded an estimated hazard ratio (HR) of 4.00 (1.53, 10.44) in comparing high risk membership to low risk; for intermediate risk membership versus low risk, the modeling yielded HR = 2.29 (0.98, 5.33). Other methods were also investigated. All methods for generating predictors of cognitive transition yielded statistically significant likelihood ratio statistics but modest concordance statistics. We concluded that the posterior probabilities from normal mixture modeling allow for risk stratification that is data-driven and modestly predictive of later cognitive decline. Incorporating other covariates may enhance predictions.

Steve Sain Jupiter Intel (Invited Presentation)

Data science and assessing risk in a changing climate. There is a long history of research and development at the intersection of applied statistics, machine learning, and climate science that has led to 1) improvements in our understanding of the Earth's climate and how that climate is changing, 2) advances in climate modeling and the use of climate model output, and 3) assessments of the impacts of climate change. The study of the impacts of climate change has provided the foundation for the emerging area of climate risk analytics, and data science is playing a key role in quantifying the impact of a changing climate on perils such as flood, heat, and fire. In this talk, I will briefly discuss some background on data science and climate science and highlight some current research areas at their intersection. In addition, I'll present an overview of climate risk analytics, focusing on research areas such as extremes, emulators, and downscaling. I'll also provide some examples of how companies and other organizations are using climate risk analytics to help assess and manage climate change-related risk.

Cadet Allie Griffith & Cadet Michael Stanis , United States Air Force Academy

ChatGPT and increased learning experience. Discover the potential of ChatGPT (a large language model that has gained popularity for its ability to naturally and intelligibly chat with users) and learn how it can enhance the learning experience at an undergraduate university. While the conversation surrounding ChatGPT in education has been heavily focused on cheating, this presentation will showcase the exciting possibilities ChatGPT brings to personalized education, on-demand tutoring, critical thinking development, and interactive problem solving. Join us to explore tangible ways that students can utilize ChatGPT to improve academics and create an engaging learning environment.

Erika Esquinca, CU Anschutz Medical Campus Colorado School of Public Health

METABOLOMICS PREDICTION FOR GENETIC DISCOVERY IN COPD USING METABOXCAN
PrediXcan is a well-known gene expression prediction workflow designed to exploit the genetic control of phenotype through gene regulation to identify trait associated genes. Novel extensions to PrediXcan seek to exploit the genetic control of other molecular measures for prediction, including metabolites. The purpose of this thesis is to develop and assess prediction models of metabolites from single nucleotide polymorphisms (SNPs) to enable broader investigation of metabolomics in studies with genome-wide SNP data. We aimed to validate novel MetaboXcan models derived from the GUARDIAN population in the COPDGene study, and to derive our own COPDGene specific models to assess the impact of cross-population differences in two machine learning methods (Bayesian Sparse Linear Mixed Models vs. Nested Cross Validated Elastic Net Models) on metabolite prediction accuracy. We developed reproducible workflows and tools to perform prediction analyses on the cloud based NHLBI BioData Catalyst Ecosystem. Of 1,272 metabolites predicted from GUARDIAN SNPs in 6,760 COPDGene subjects, prediction accuracy was relatively low with only 220 significant correlations between predicted and observed metabolite levels, and only 24 metabolites with prediction accuracy greater than 0.1. When comparing prediction accuracy of GUARDIAN models with models trained in COPDGene, we found 719 metabolite models produced a

significant Pearson correlation between predicted and observed, and 272 metabolites with prediction accuracy greater than 0.1. Findings suggest that source populations used to train the models will affect prediction accuracy. When assessing the two different methods utilized to create prediction models, the Bayesian framework produced substantially better models. Bayesian models had an average of 745 SNPs used per model compared to the Elastic Net methods which had an average of 51 SNPs used per model. This can be attributed to the Bayesian methods ability to capture small effect sizes, therefore obtaining more SNPs per model, and improving prediction accuracy. However, further investigation into alternative analytic methods could improve metabolite prediction accuracy.

Troy Wixson, Colorado State University

Modeling Wildfire Risk Wildfire risk is greatest during high winds after sustained periods of dry and hot conditions. This paper is an extreme event risk attribution study which aims to usefully answer whether extreme wildfire seasons are more likely now than under past climate. This requires modeling temporal dependence at extreme levels. We propose the use of transformed-linear time series models which are constructed similarly to traditional ARMA models while having a dependence structure that is tied to a widely used framework for extremes (regular variation). We fit the models to the extreme values of the seasonally adjusted Fire Weather Index (FWI) time series to capture the dependence in the upper tail for past and present climate. Ten-thousand fire seasons are simulated from each fitted model and we compare the proportion of simulated high-risk fire seasons to quantify the increase in risk. Our method suggests that the risk of experiencing an extreme wildfire season in Grand Lake, Colorado under current climate has increased dramatically compared to the risk under the climate of the mid-20th century. Our method also finds some evidence of increased risk of extreme wildfire seasons in Quincy, California, but large uncertainties do not allow us to reject a null hypothesis of no change.

Mantautas Rimkus, Colorado State University

Classification methods for fault localization in future power grids The evolution of power flow structures in transmission grids, driven by the rapid growth of renewable energy generation and bidirectional power flows, is expected to bring significant changes in the coming years. This poses a challenge for traditional fault localization methods, especially in scenarios where partial observability limits the accuracy of fault localization. While classification methods have been proposed for fault localization, their effectiveness depends on the availability of labeled data, which is often not practical in real-life situations. Therefore, the objective of our study is to bridge the gap between partial and full observability of the power grid, and develop fault localization methods that can operate efficiently even in scenarios with limited PMU availability. We propose using Graph Neural Networks with combination of statistical fault localization methods to solve the forementioned problems. Our contribution to the field of fault localization aims to enable the adoption of effective fault localization methods for future power grids

Danielle Dematteis, CSU Statistics Department

Distributed lag interaction model for modeling maternal exposure to air pollution
Maternal exposure to air pollution during pregnancy has a substantial public heath impact. Epidemiological evidence supports an association between maternal exposure to air pollution and low birth weight. A popular method to estimate the linear association between maternal exposure to air pollution and birth weight is a distributed lag model (DLM), which regresses birth weight onto maternal exposure history observed at multiple time points during pregnancy. However, the standard DLM framework does not allow linear associations of exposure on birth weight to vary for each individual based on a continuous variable. We propose a distributed lag interaction model that allows modification of the exposure-time-response associations across individuals by including an interaction between a continuous modifying variable and the exposure history. Our model framework is an extension of a DLM that uses a cross-basis, or bi-dimensional function space, to simultaneously describe both the modification of the exposure-response relationship and its temporal structure. Through simulations, we showed that our model with penalization out-performs a standard DLM when the true exposure-time-response associations depend on a continuous covariate. Using a Colorado USA birth cohort, we estimated the association between birth weight and PM2.5 modified by an area-level metric of health and social adversities from Colorado EnviroScreen.

Connor Gibbs, Colorado State University

ECoHeN: A Hypothesis Testing Framework for Extracting Communities from Heterogeneous Networks. Community discovery is a process of identifying assortative communities in a network: collections of nodes which are densely connected within but sparsely connected to the rest of the network. While community discovery has been extensively studied, there are few techniques available for heterogeneous networks that contain different types of nodes and possibly different connectivity patterns between the node types. In this talk, we introduce a framework called ECoHeN to extract communities from a heterogeneous network in a statistically meaningful way. ECoHeN uses a heterogeneous configuration model as a reference distribution to identify communities that are significantly more densely connected than expected given the node types and connectivity of its membership. The ECoHeN algorithm extracts communities one at a time using a dynamic set of iterative updating rules, is guaranteed to converge, and imposes no constraints on the type composition of extracted communities. To our knowledge, ECoHeN is the first method that can distinguish and identify both homogeneous and heterogeneous, possibly overlapping, community structure in a network. We demonstrate the utility of ECoHeN in the context of a popular political blogs network to identify collections of blogs that reference one another more than expected considering the ideology of its members.

https://arxiv.org/abs/2212.10513"