MSSISS 2022

MARCH 10–11

Michigan Student Symposium for Interdisciplinary Statistical Science

Ann Arbor, MI, USA

PROGRAM

TABLE OF CONTENTS

ABOUT

The *Michigan Student Symposium for Interdisciplinary Statistical Sciences* (MSSISS) is an annual event organized by graduate students in the Biostatistics, Electrical Engineering & Computer Science (EECS), Industrial & Operations Engineering (IOE), Statistics and Survey and Data Science (MPSDS) departments at the University of Michigan.

The goal of this symposium is to create an environment that allows communication across related fields of statistical sciences and promotes interdisciplinary research among graduate students and faculty. It encourages graduate students to present their work, share insights, and exposes them to diverse applications of statistical sciences. Though hosted by five departments we extend our invitation to graduate students from all departments across the University to present their statistical research in the form of an oral paper presentation or a poster presentation. It also provides an excellent environment for interacting with students and faculty from other areas of statistical research on campus.

MSSISS is an opportunity for interdisciplinary research and discussion across the fields of statistical sciences. Calling all graduate students (as well as talented undergraduates)! Come along, present your work, share insights, and learn about the diverse applications of statistical sciences.

MARCH 10TH

9:00am–12:00pm Concourse	**Registration**
10:00am–10:15am Vandenberg	**Opening remarks**
10:15am–12:00pm Vandenberg	**Session I**
1:00pm–2:30pm Vandenberg	**Speed Session**
2:30pm–4:00pm Hussey	**Poster Session**
4:00pm–5:00pm Vandenberg	**Michigan Speaker - Vijay Subramanian**

MARCH 11TH

8:30am–10:00am Vandenberg	**Session II**
10:00am–10:30am Ballroom	**Breakfast**
10:30am–12:00pm Vandenberg	**Session III**
12:00pm–1:00pm Ballroom	**Lunch**
1:00pm–2:30pm Vandenberg	**Session IV**
3:00pm–4:00pm Ballroom	**Keynote Speaker - Bin Yu**
4:00pm–4:30pm Ballroom	**Awards & Closing Remarks**

KEYNOTE SPEAKER

Bin Yu

Professor of Statistics and EECS
University of California, Berkeley

Bin Yu is Chancellor's Distinguished Professor and Class of 1936 Second Chair in the departments of statistics and EECS at UC Berkeley. She leads the Yu Group which consists of students and postdocs from Statistics and EECS. She was formally trained as a statistician, but her research extends beyond the realm of statistics. Together with her group, her work has leveraged new computational developments to solve important scientific problems by combining novel statistical machine learning approaches with the domain expertise of her many collaborators in neuroscience, genomics and precision medicine. She and her team develop relevant theory to understand random forests and deep learning for insight into and guidance for practice.

She is a member of the U.S. National Academy of Sciences and of the American Academy of Arts and Sciences. She is Past President of the Institute of Mathematical Statistics (IMS), Guggenheim Fellow, Tukey Memorial Lecturer of the Bernoulli Society, Rietz Lecturer of IMS, and a COPSS E. L. Scott prize winner. She holds an Honorary Doctorate from The University of Lausanne (UNIL), Faculty of Business and Economics, in Switzerland. She has recently served on the inaugural scientific advisory committee of the UK Turing Institute for Data Science and AI, and is serving on the editorial board of Proceedings of National Academy of Sciences (PNAS)

Interpreting Deep Neural Networks towards Trustworthiness

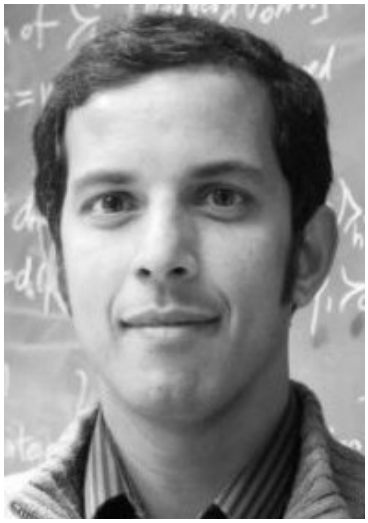
Friday, March 11th, 3pm

Ballroom

Recent deep learning models have achieved impressive predictive performance by learning complex functions of many variables, often at the cost of interpretability. In this talk, I will begin with a discussion on defining interpretable machine learning in general and describe our agglomerative contextual decomposition (ACD) method to interpret neural networks. ACD attributes importance to features and feature interactions for individual predictions and has brought insights to NLP/computer vision problems and can be used to directly improve generalization in interesting ways.

I will then focus on scientific interpretable machine learning. Building on ACD's extension to the scientifically meaningful frequency domain, an adaptive wavelet distillation (AWD) interpretation method is developed. AWD is shown to be both outperforming deep neural networks and interpretable in two prediction problems from cosmology and cell biology. Finally, I will address the need to quality-control the entire data science life cycle to build any model for trustworthy interpretation and introduce our Predictabiltiy Computability Stability (PCS) framework for such a data science life cycle...

MICHIGAN SPEAKER



Vijay Subramanian

Associate Professor of EECS
University of Michigan

Vijay Subramanian is an Associate Professor in the EECS Department at the University of Michigan since 2014. After graduating with his Ph.D. from UIUC in 1999, he did a few stints in industry, research institutes and universities in the US and Europe before his current position. His main research interests are in stochastic modeling, networks, applied probability and applied mathematics. A large portion of his past work has been on probabilistic analysis of communication networks, especially analysis of scheduling and routing algorithms. In the past he has also done some work with applications in information theory, mathematical immunology and coding of stochastic processes. His current research interests are on game theoretic and economic modeling of socio-technological systems and networks, reinforcement learning and the analysis of associated stochastic processes.

Statistics in Action for Reinforcement Learning

Thursday, March 10th, 4pm

Vandenberg

The use of sufficient statistics for learning, and parametric and non-parametric learning of models is common in statistics. In this talk we highlight the relevance and use of these methodologies in reinforcement learning, i.e., data-driven control of unknown stochastic dynamical systems.

We start by describing the use of sufficient statistics, particularly the notion of an information state and its relaxation to an approximate information state, to develop criteria for simpler representations of decentralized multi-agents systems that can be use to obtain close-to-optimal policies in a data-driven setting. Specifically, within the centralized training with distributed execution framework, we develop conditions that an approximate information state based simpler representation should satisfy so that low regret policies can be obtained. A key new feature that we highlight is the compression of actions of the agents (and accompanying contribution to regret) that occurs as a result of an approximate information state. This is joint work with Hsu Kao at the University of Michigan that is to be presented at AISTATS 2022.

If time permits, we will also discuss recent work on challenges in parametric and non-parametric learning based optimal control in stochastic dynamic systems using self-tuning adaptive control ideas. In this work, we study learning-based optimal admission control for a classical Erlang-B blocking system with unknown service rate, i.e., an M/M/k/k queueing system. At every job arrival, a dispatcher decides to assign the job to an available server or to block it. Every served job yields a fixed reward for the dispatcher, but it also results in a cost per unit time of service. Our goal is to design a dispatching policy that maximizes the long-term average reward for the dispatcher based on observing the arrival times and the state of the system at each arrival; critically, the dispatcher observes neither the service times nor departure times. We develop an asymptotically optimal learning based admission control policy and use it to show how the extreme contrast in the certainty equivalent optimal control policies in our problem leads to difficulties in learning; these are illustrated using our regret bounds for different parameter regimes. This is joint work with Saghar Adler at the University of Michigan and Mehrdad Moharrami at the University of Illinois at Urbana-Champaign.

SESSION I

Moderated by

Curstiss Engstrom

PhD Student

Program in Survey and Data Science

March 10th, 10:15am – 12:00pm

Vandenberg

Timothy Baker

Fifth-year PhD Student
EECS

Collaborator

John Hayes

Leveraging Correlation to Improve Accuracy in Stochastic Computing

In stochastic computing, streams of random bits are used to perform low-cost computation. For example, two random bitstreams can be multiplied using a single AND gate whereas conventional digital multipliers require hundreds of logic gates. Efficient multiplication has made stochastic computing a promising design paradigm for low-cost hardware implementations of digital filters, image processing algorithms and neural networks. However, due to their inherant randomness, stochastic circuits yield approximate computation results and have a fundamental accuracy-latency tradeoff. Sometimes this trade-off is poor and a stochastic circuit will require high latency to reach practical accuracy thresholds. Surprisingly, our recent work shows how correlation can be leveraged to drastically improve the accuracy of some important stochastic circuit designs and ultimately lower their latency. We introduce two techniques, full correlation and precise sampling, which improve the accuracy of multiplexer-based random bitstream adders by 4x to 16x while reducing the circuit area by about 35%. This accuracy improvement translates into a significantly lower required latency as demonstrated by a digital filtering case study.

Dan Kessler

Fifth-year PhD Student
Statistics

Collaborator

Elizaveta Levina

Inference for Canonical Directions in Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a method for analyzing pairs of random vectors; it learns a sequence of paired linear transformations such that the resultant canonical variates are maximally correlated within pairs while uncorrelated across pairs. The parameters estimated by CCA include both the “canonical correlations” as well as the “canonical directions” which characterize the transformations. CCA has seen a resurgence of popularity with applications including brain imaging and genomics where the goal is often to identify relationships between high-dimensional -omics data with more moderately sized behavioral or phenotypic measurements. Inference in CCA applications is typically limited to testing whether the canonical correlations are nonzero. Inference for the canonical directions has received relatively little attention in the statistical literature and in practice the directions are interpreted descriptively. We discuss several approaches for conducting inference on canonical directions obtained by CCA. We conduct thorough simulation studies to assess inferential validity in various settings and apply the methods to a brain imaging data set.

Lulu Shang
Fourth-year PhD Student
Biostatistics

Collaborator
Xiang Zhou

Spatially Aware Dimension Reduction for Spatial Transcriptomics

Spatial transcriptomics are a collection of genomic technologies that have enabled transcriptomic profiling on tissues with spatial localization information. Analyzing spatial transcriptomic data is computationally challenging, as the data collected from various spatial transcriptomic technologies are often noisy and display substantial spatial correlation across tissue locations. Here, we develop a spatially-aware dimension reduction method, SpatialPCA, that can extract a low dimensional representation of the spatial transcriptomics data with enriched biological signal and preserved spatial correlation structure, thus unlocking many existing computational tools previously developed in single-cell RNAseq studies for tailored and novel analysis of spatial transcriptomics. We illustrate the benefits of SpatialPCA for spatial domain detection and explores its utility for trajectory inference on the tissue and for high-resolution spatial map construction. In the real data applications, SpatialPCA identifies key molecular and immunological signatures in a newly detected tumor surrounding microenvironment, including a tertiary lymphoid structure that shapes the gradual transcriptomic transition during tumorigenesis and metastasis. In addition, SpatialPCA detects the past neuronal developmental history that underlies the current transcriptomic landscape across tissue locations in the cortex.

Yutong Wang
Sixth-year PhD Student
EECS

Collaborator
Clayton Scott

VC dimension of partially quantized neural networks in the overparametrized regime

Vapnik-Chervonenkis (VC) theory has so far been unable to explain the small generalization error of overparametrized neural networks. Indeed, existing applications of VC theory to large networks obtain upper bounds on VC dimension that are proportional to the number of weights, and for a large class of networks, these upper bound are known to be tight. In this work, we focus on a class of partially quantized networks that we refer to as hyperplane arrangement neural networks (HANNs). Using a sample compression analysis, we show that HANNs can have VC dimension significantly smaller than the number of weights, while being highly expressive. In particular, empirical risk minimization over HANNs in the overparametrized regime achieves the minimax rate for classification with Lipschitz posterior class probability. We further demonstrate the expressivity of HANNs empirically. On a panel of 121 UCI datasets, overparametrized HANNs are able to match the performance of state-of-the-art full-precision models.

Yuqi Zhai
Fourth-year PhD Student
Biostatistics

Collaborator
Peisong Han

Data Integration with Oracle Use of External Information from Heterogeneous Populations

It is common to have access to summary information from external studies. Such information can be useful in model fitting for an internal study of interest and can improve parameter estimation efficiency when incorporated. However, external studies may target populations different from the internal study, in which case an incorporation of the corresponding information may introduce estimation bias. We develop a penalized constrained maximum likelihood (PCML) method that simultaneously achieves (i) selecting the external studies whose target populations match the internal study's such that their information is useful for internal model fitting, and (ii) automatically incorporating the corresponding information into internal estimation. The PCML estimator has the same efficiency as an oracle estimator that knows which external information is useful and fully incorporates that information alone. A detailed theoretical investigation is carried out to establish asymptotic properties of the PCML estimator, including estimation consistency, parametric rate of convergence, external information selection consistency, asymptotic normality, and oracle efficiency. An algorithm for numerical implementation is provided, together with a data-adaptive procedure for tuning parameter selection. Numerical performance is investigated through simulation studies and an application to a prostate cancer study is also conducted.

SESSION II

Moderated by

Alexander Ritchie

PhD Student

EECS

March 11th, 8:30am – 10:00am

Vandenberg

Trong Dat Do

Third-year PhD Student
Statistics

Collaborators

Jiacheng Zhu

Aritra Guha

XuanLong Nguyen

Ding Zhao

Mengdi Xu

Functional Optimal Transport: map estimation and domain adaptation for functional data

We introduce a formulation of optimal transport problem for distributions on function spaces, where the stochastic map between functional domains can be partially represented in terms of an (infinite-dimensional) Hilbert-Schmidt operator mapping a Hilbert space of functions to another. For numerous machine learning tasks, data can be naturally viewed as samples drawn from spaces of functions, such as curves and surfaces, in high dimensions. Optimal transport for functional data analysis provides a useful framework of treatment for such domains. To this end, we develop an efficient algorithm for finding the stochastic transport map between functional domains and provide theoretical guarantees on the existence, uniqueness, and consistency of our estimate for the Hilbert-Schmidt operator. We validate our method on synthetic datasets and examine the geometric properties of the transport map. Experiments on real-world datasets of robot arm trajectories further demonstrate the effectiveness of our method on applications in domain adaptation.

Jinming Li

Third-year PhD Student
Statistics

Collaborators

Gongjun Xu

Ji Zhu

Network Latent Space Model with Hyperbolic Geometry

Network data are prevalent in various scientific and engineering fields, including sociology, economics, neuroscience, and so on. While latent space models are widely used in analyzing network data, the geometric effect of latent space remains an important but unsolved problem. In this work, we propose a hyperbolic network latent space model with a learnable curvature parameter, which allows the proposed model to fit network data with the most suitable latent space. We theoretically justify that learning the optimal curvature is essential to minimize the embedding error for all hyperbolic embedding methods beyond network latent space models. We also establish consistency rates for maximum-likelihood estimators and develop an estimation approach with manifold gradient optimization, both of which are technically challenging due to the non-linearity and non-convexity of hyperbolic distance metric. We further illustrate the superiority of the proposed model and the geometric effect of latent space with extensive simulation studies followed by a Facebook friendship network application.world datasets of robot arm trajectories further demonstrate the effectiveness of our method on applications in domain adaptation.

Jieru Shi
Second-year PhD Student
Biostatistics

Collaborators
Zhenke Wu
Walter Dempsey

Assessing Time-Varying Causal Effect Moderation in the Presence of Cluster-Level Treatment Effect Heterogeneity

The micro-randomized trial (MRT) is a sequential randomized experimental design to empirically evaluate the effectiveness of mobile health (mHealth) intervention components that may be delivered at hundreds or thousands of decision points. MRTs have motivated a new class of causal estimands, termed "causal excursion effects", for which semiparametric inference can be conducted via a weighted, centered least-squares criterion (Boruvka et al., 2018). Existing methods assume between-subject independence and non-interference. Deviations from these assumptions often occur. In this paper, causal excursion effects are revisited under potential cluster-level treatment effect heterogeneity and interference, where the treatment effect of interest may depend on cluster-level moderators. The utility of the proposed methods is shown by analyzing data from a multi-institution cohort of first-year medical residents in the United States.

Natasha Stewart
Third-year PhD Student
Statistics

Collaborators
Snigdha Panigrahi
Elizaveta Levina

Post-Selection Inference for Multitask Regression with Shared Sparsity

With the growing complexity of modern data, it is increasingly common to select a data model only after performing an exploratory analysis. The field of selective inference has arisen to provide valid inference following the selection of a model through such data-adaptive procedures. The contribution of this work is to develop post-selection inference tools for multitask learning problems. Multitask learning is used to model a set of related response variables from the same set of features, improving predictive performance relative to methods that handle each response variable separately. Ignoring the shared structure for the sake of obtaining valid inference would come at a significant cost in terms of power, and thus new methods are needed. Motivated by applications in neuroimaging, we consider problems where several response variables must each be modeled using some sparse subset of the shared features. This setup can arise, for instance, when a series of related phenotypes are modeled as a function of brain imaging data.

We propose a two-stage protocol for joint model selection and inference. In stage one, we adapt a penalty approximation to jointly identify the relevant covariates for each task, proceeding to fit a series of linear model using the selected features. In stage two, a new conditional approach is proposed to infer about the selected models, utilizing a refinement of the selection event. An approximate system of estimating equations for maximum likelihood inference is developed that can be solved via a single convex optimization problem. This enables us to efficiently form confidence intervals with roughly the desired coverage probability through MLE-based inference. We test our two-stage procedure on simulated data, demonstrating that our methods yield tighter confidence intervals than alternatives such as data splitting. Finally, we consider an application in neuroscience involving high-dimensional fMRI data and several related cognitive tasks.

Jung Yeon Won
Fifth-year PhD Student
Biostatistics

Collaborators
Michael R. Elliott
Brisa N. Sanchez

Integrating food environment exposures from multiple longitudinal databases

The majority of built environment health studies rely on secondary sources to enumerate local food environments and conduct analyses to contextualize population health and health behaviors within a neighborhood's retail environments. Such secondary commercial databases often provide longitudinal point-referenced data, which enables longitudinal studies that characterize health outcomes in relation to the dynamic food environment. However, there are concerns about measurement error when quantifying environmental influences with longitudinal secondary commercial sources due to the incompleteness of listings. To alleviate the ascertainment error problem, combining multiple databases can be a promising strategy in particular for time-varying exposures as field validation is not feasible for historical exposure measures. Given the quality scores of each database, we propose a method that incorporates source quality to integrate conflicting time-varying exposures that are from different data sources. To model the latent time-varying count exposure, we extend the Poisson INAR(1) model and take a Bayesian nonparametric approach to flexibly discover clusters of location-specific time series of exposures. By resolving the discordance between different databases, our method obtains an unbiased health effect of unobservable series of true exposures.

SESSION III

Moderated by

Lap Sum Chan

PhD Student

Biostatistics

March 11th, 10:30am – 12:00pm

Vandenberg

Rupam Bhattacharyya
Fourth-year PhD Student
Biostatistics

Collaborators
Nicholas Henderson
V. Baladandauythapani

fiBAG: Functional Integrative Bayesian Analysis of High-dimensional Multiplatform Genomic Data

Large-scale multi-omics datasets offer complementary, partly independent, high-resolution views of the human genome. Modeling and inference using such data poses challenges like high-dimensionality and structured dependencies but offers potential for understanding the complex biological processes characterizing a disease. We propose fiBAG, an integrative hierarchical Bayesian framework for modeling the fundamental biological relationships underlying such cross-platform molecular features. Using Gaussian processes, fiBAG identifies mechanistic evidence for covariates from corresponding upstream information. Such evidence, mapped to prior inclusion probabilities, informs a calibrated Bayesian variable selection (cBVS) model identifying genes/proteins associated with the outcome. Simulation studies illustrate that cBVS has higher power to detect disease-related markers than non-integrative approaches. A pan-cancer analysis of 14 TCGA cancer datasets is performed to identify markers associated with cancer stemness and patient survival. Our findings include both known associations like the role of RPS6KA1/p90RSK in gynecological cancers and interesting novelties like EGFR in gastrointestinal cancers.

Kyle Gilman
Fifth-year PhD Student
EECS

Collaborators
David Hong
Laura Balzano
Jeffrey Fessler

Streaming Probabilistic PCA for Missing Data with Heteroscedastic Noise

Streaming principal component analysis (PCA) has been an integral tool in large-scale machine learning for rapidly estimating low-dimensional subspaces of very high dimensional and high arrival-rate data with missing entries and corrupting noise. However, modern trends increasingly combine data from a variety of sources, meaning they may exhibit heterogeneous quality across samples. Since standard streaming PCA algorithms do not account for non-uniform noise, their subspace estimates quickly degrade. On the other hand, recently proposed heteroscedastic probabilistic PCA (HPPCA) is limited in its practicality since it does not handle missing entries and streaming data, nor can it adapt to non-stationary behavior in time series data. In this work, we propose the Streaming HeteroscedASTic Algorithm for PCA (SHASTA-PCA) to bridge this divide. Our method uses a stochastic alternating expectation maximization approach to jointly learn the low-rank latent factors and unknown noise variances from streaming data with missing entries and heteroscedastic noise, all while maintaining a low memory and computational footprint. Numerical experiments validate the superior performance of our method compared to state-of-the-art streaming PCA algorithms.

Mao Li
First-year PhD Student
Program in Survey and Data Science

Using network analysis and clustering to evaluate the effectiveness of the US Census Bureau's social media campaign about self-completing the 2020 census

From the start of data collection for the 2020 US Census, official and celebrity users tweeted about the importance of everyone being counted in the census and urged followers to complete the questionnaire. At the same time, skeptical social media posts about the census became increasingly common. This study aims to identify and investigate the influence of Twitter user communities on self-completion rate, according to Census Bureau data, for the 2020 Census. Using a network analysis method, Community Detection, and a clustering algorithm, Latent Dirichlet Allocation, three prototypical users were identified: "official" (i.e., government agency), "promoting-census," and "census skeptics" users. The census skeptics group was motivated by events and speeches about which an influential person had tweeted and became the largest community over the study period. The promoting-census community was less motivated by specific events and was consistently more active than the census skeptics community. The official user community was the smallest of the three, but their messages seemed to have been amplified by promoting-census celebrities and politicians. We found that the daily size of the promoting-census users group – but not the other two – predicted the Census 2020 Internet self-completion rate within 3 days after a tweet was posted, suggesting that the census social media campaign was successful apparently due to the help of promoting-census celebrities, who encouraged people to fill out the census amplifying official user tweets. This finding demonstrates that a social media campaign can positively affect public behavior about an essential national project like the decennial census.

Robert Lunde
Post-doctoral Fellow
Statistics

Collaborators
Elizaveta Levina
Ji Zhu

Conformal Prediction for Network Regression

An important problem in network analysis is predicting a node attribute using nodal covariates and summary statistics computed from the network, such as graph embeddings or local subgraph counts. While standard regression methods may be used for prediction, statistical inference is complicated by the fact that the nodal summary statistics often exhibit a nonstandard dependence structure. When the underlying network is generated by a graphon, we show that conformal prediction methods are finite-sample valid under a very mild condition on the network summary statistics. We also prove that a form of asymptotic conditional validity is achievable using standard nonparametric regression methods.

Jing Ouyang
Third-year PhD Student
Statistics

Collaborators
Kean Ming Tan
Gongjun Xu

High-dimensional inference on Generalized Linear Models with Unmeasured Confounders

In the high-dimensional setting, the inference problems on the relationship between the response and the covariates are extensively studied for their wide applications in medicine, economics, and many other fields. In many applications, the covariates are often associated with unmeasured confounders such as in studying the genetic effect on a certain disease, the gene expressions are confounded by some unmeasured environmental factors. In this case, the standard methods may fail due to the existence of the unmeasured confounders. Recent studies address this problem in the context of linear models whereas the problem in generalized linear framework is less investigated. In this paper, we consider a generalized linear framework and propose a debiasing approach to address this high-dimensional problem, while adjusting for the effect of unmeasured confounders. We establish the asymptotic distribution for the debiased estimator. A simulation study and an application of our method on a genetic data set are performed to demonstrate the validity of this approach.

SESSION IV

Moderated by

Cheoljoon Jeong

PhD Student

IOE

March 11th, 1:00pm – 2:30pm

Vandenberg

Derek Hansen

Fourth-year PhD Student
Statistics

Collaborators

Ismael Mendoza

Runjing Liu

Jeffrey Regier

Scalable Bayesian Inference for Detecting and Deblending Stars and Galaxies in Crowded Fields

In images from astronomical surveys, astronomical objects such as stars and galaxies often overlap. Deblending is the task of identifying and characterizing the individual light sources that make up such images. We propose the Bayesian Light Source Separator (BLISS), which enables the detection, characterization, and reconstruction of individual stars and galaxies. BLISS posits a fully generative model of an astronomical image and its associated catalog, which consists of locations, brightness, classification (star or galaxy), and the galaxy shape.

First, to learn a distribution of galaxy shapes, we train a Variational Autoencoder (VAE) on simulated images of single galaxies. The VAE works by associating each galaxy with a low-dimensional latent representation from which all relevant information about its shape can be reconstructed. Then, to efficiently sample from the posterior distribution of the catalog given the image, we use amortized Variational Inference (VI) via a flexible neural network encoder. Our encoder consists of three stages. First, we sample the number of objects in the image and their locations conditional on the image. Then, we calculate the probability each object is a galaxy or star and sample the label. Finally, for each labeled galaxy, we sample the associated latent representation. Using the VAE, these latent representations can be reconstructed into individual galaxies, enabling downstream astronomical tasks that rely on the deblended morphology. Unlike traditional VI, the encoder is trained by alternating between the forward Kullback-Liebler (KL) divergence using simulated images and the reverse KL divergence using real images. Using the Sloan Digital Sky Survey (SDSS) dataset, we demonstrate that BLISS can find, classify, and reconstruct stars and galaxies identified in previous surveys with both high recall and precision.

Ai Rene Ong

Fifth-year PhD Student
Program in Survey and Data Science

Collaborators

Sunghee Lee

Michael Elliott

Respondent Driven Sampling Design Considerations

Respondent Driven Sampling (RDS) has been used as a method to sample hard-to-sample populations, leveraging the social networks of the initial respondents, typically selected through convenience sampling, to reach more people from the target population. Respondents are asked to invite their eligible peers to participate in the study, and this process continues until the sample size is reached. Although there have been some general recommendations for RDS best practices (e.g., conducting formative studies, a small number of seed respondents), efforts to study the contributions of these design decisions on the productivity of RDS peer recruitment have been hindered by incomplete reporting of RDS methodology in the literature. This study presents an exploratory analysis of the associations of various RDS design decisions on peer recruitment productivity. The data used is from a survey of researchers who have published an article using RDS or have grants funded for research using RDS from 2009 to 2020. These researchers were sampled from a database that represents a census of RDS researchers. A hundred and twenty-one researchers completed the survey which asked about the design of their RDS data collection. Preliminary results indicate that fielding an RDS survey on the web is associated with better productivity, and this effect is moderated by the type of population the RDS study is targeting. Giving more than one form of instructions for peer recruitment appeared to help with peer recruitment productivity. However, conducting formative research prior to data collection was not associated with peer recruitment productivity.

Yifan Hu
First-year Master's Student
Statistics

Estimating An Optimal Individualized Treatment Rule for Guiding the Initial Treatment Decision on Child/Adolescent Anxiety Disorder

Collaborators
Tuo Wang
Scott N. Compton
Daniel Almirall

Designing an Individualized Treatment Rule (ITR) to guide clinicians on deciding personalized treatment plans for patients is an important research goal in treating child/adolescent anxiety disorder. An ITR is a special case of a dynamic treatment regimen when there is a single decision rule. In this research, an ITR is said to be optimal if it maximizes the expectation of a pre- specified clinical outcome when used to assign treatment to a population of interest. Our goal is to establish and to evaluate an optimal IRT, which guides the decision among sertraline (SRT), cognitive behavior therapy (CBT), and their combination (COMB) as the initial treatment for children or adolescents with anxiety referring to Child/Adolescent Anxiety Multimodal Study (CAMS). The Study (CAMS) is a completed federally-funded, multi-site, randomized placebo-controlled trial that examined the relative efficacy of cognitive-behavior therapy (CBT), sertraline (SRT), and their combination (COMB) against pill placebo (PBO) for the treatment of separation anxiety disorder (SAD), generalized anxiety disorder (GAD) and social phobia (SOP) in children and adolescents. Based on the Pediatric Anxiety Rating Scale (PARS) outcome from 412 participants randomly assigned to CBT (139), SRT (133), or COMB (140), we propose a four-step technique, to estimate an optimal ITR using the CAMS data that leads to the minimal symptoms, on average. The four steps are: (1) Split the data for training (70%) and evaluation (30%). In the training data set: (2) Prune the baseline covariates according to their contribution level to model with a specified variable selection algorithm for subset analysis; (3) Use a novel method called "Decision List" to create two prospective interpretable and parsimonious IRTs. (4) Use the test data to evaluate the two candidate IRTs versus providing SRT only, CBT only, or COMB to participants with PARS results.

Peter MacDonald
Fourth-year PhD Student
Statistics

Continuous-time latent process network models

Collaborators
Elizaveta Levina
Ji Zhu

Network data are often collected through the observation of a complex system over time, leading to time-stamped network snapshots. Methods in statistical network analysis are traditionally designed for a single network, but applying these methods to a time-aggregated network can miss important temporal structure in the data. In this work, we provide an approach to estimating the expected network in continuous time. We parameterize the network expectation through time-varying positions, such that the activity of each node is governed by a low-dimensional latent process. To tractably estimate these processes, we assume their components come from a fixed, finite-dimensional function basis. We provide a gradient descent estimation approach, establish theoretical results for its convergence, compare our method to competitors, and apply it to a real dynamic network of international political interactions.

Lam Tran
Fourth-year PhD Student
Biostatistics

A fast algorithm for fitting the constrained lasso

Collaborators
Hui Jiang
Gen Li

Background: The constrained lasso is a flexible framework that augments the lasso by allowing for imposition of additional structure on regression coefficients. Despite the constrained lasso's broad utility in compositional microbiome analysis and gene pair discovery, among many other applications, current methods for fitting the constrained lasso do not computationally scale and are limited to linear and logistic models with only simple constraint structures. No existing methods deal with survival data, limiting the range of potential clinical applications.
Methods: We proposed a novel approach for fitting the constrained lasso by leveraging candidate covariate subsets of increasing size from the unconstrained lasso in an efficient alternating direction method of multipliers algorithm. We found that using this approach can accelerate the convergence of the constrained lasso. We tested the ability of our method to quickly fit the constrained lasso with simulated and real-world data types under a variety of constraint structures.
Results: Our proposed algorithm led substantial speedups in solving the regularization path of the constrained lasso for simulated data, even in complex cases where not all predictors are penalized and constrained equally. The utility and speed of our method were maintained when we considered two real-world data examples: a compositional microbiome dataset for binary periodontal disease status and a microarray dataset for multiple myeloma survival, neither of which could be solved when the constrained lasso is naively fit on the full set of predictors.
Significance: Our proposed algorithm dramatically reduces the time required to fit the constrained lasso even in real-data settings with complex constraint structures. Our computationally inexpensive approach increases the range of potential applications for the flexible and robust constrained lasso, being able to quickly perform variable selection with multiple response types and constraints.

SPEED SESSION

Moderated by

Simon Fontaine

PhD Student
Statistics

March 10th, 1:00pm – 2:30pm

Vandenberg

Some speed presenters also participate in the Poster Session; they can be identified by their poster number at the top right of the abstract.

Madeline Abbott

Second-year PhD Student
Biostatistics

Collaborators

Walter Dempsey
Inbal Nahum-Shani
Jeremy Taylor

Modeling cigarette use with mobile health data from a study on smoking cessation

1

Ecological momentary assessment (EMA), which consists of frequently delivered surveys sent to subjects' smartphones, allows for the collection of data in real time and in subjects' natural environments. As a result, data collected using EMA can be particularly useful in understanding rapid temporal variation in subjects' states and environments, as well as how these states and environments relate to outcomes of interest. Focusing on a study of smoking cessation, we use data collected via EMA from current smokers who recently quit to model changes in their emotional state over time and to assess possible associations between their state and risk of cigarette use. We apply a factor model to summarize 23 different emotions as two key psychological concepts: positive affect and negative affect. We then use positive and negative affect, along with a measure of urge to smoke, to model lapses in smoking cessation over time. Based on currently available data, we find that positive and negative affect summarize the emotions in our data well and that time since quit is strongly associated with risk of cigarette use.

Neophytos Charalambides

Fifth-year PhD Student
EECS

Collaborators

Alfred Hero
Mert Pilanci

Resilient and Secure Distributed Matrix Inversion

A cumbersome operation in statistics, signal processing, numerical analysis and linear algebra, optimization and machine learning, is inverting large full-rank matrices. We propose a coded computing approach for recovering matrix inverse approximations. We present an approximate matrix inversion algorithm which does not require a matrix factorization, but uses a black-box least squares optimization solver as a subroutine, to give an estimate of the inverse of real full-rank matrices. We then present a distributed framework for which our algorithm can be implemented, and show how we can leverage from sparsest-balanced MDS generator matrices to devise inverse coded computing schemes. We focus on balanced Reed-Solomon codes, which are optimal in terms of computational load; and communication from the workers to the master server. We also discuss how our algorithms can be used to compute the pseudoinverse of a full-rank matrix, and how the communication can be secured from eavesdroppers. Our approach can also be utilized for exact matrix product recovery.

Alicia Dominguez

Fourth-year PhD Student
Biostatistics

Collaborators

Sebastian Zoellner

Yuhua Zhang

Bias accumulates in polygenic risk scores constructed with larger sets of markers in multiple complex traits.

2

Polygenic risk scores (PRS) are increasingly used to predict genetic risk for many complex traits. Many PRS methods, like p-value thresholding (P&T), rely on effect size estimates from genome-wide association studies (GWAS) of predominantly European (EUR) samples which can potentially bias risk estimates for non-EUR populations, especially individuals of African (AFR) ancestry. How this bias aggregates over PRS with larger sets of markers is not well understood.

Our sample (n=5,848) consists of genotyped individuals from the University of Michigan (UM) Prechter Bipolar Study and Michigan Genome Initiative. To estimate ancestry for these individuals, we performed principal component analysis on their genomic data and data (n=2,504) from the 1,000 Genome Project (1KGP). We used publicly-available GWAS summary statistics and P&T method to infer PRS for the UM and 1KGP samples for four complex traits: bipolar disorder (BD), height, schizophrenia (SCZ), and type II diabetes. For 1KGP participants, we quantify transferability of EUR-biased genetic studies to other populations by comparing PRS across populations for multiple sets of markers. For the UM sample, we evaluate the relationship between number of markers and predictive accuracy for PRS of BD and SCZ and use regression models to evaluate factors associated with PRS.

For traits like BD, PRS calculated with more markers were more predictive of affection status in our UM sample. For 1KGP, we see directional inconsistencies across different populations for several PRS but most evident in PRS constructed with more markers. Furthermore, PRS calculated with more markers were significantly associated (p<0.05) with several ancestry principal components for both samples.

There is a tension between having more complex, informative PRS and their susceptibility to population structure. Consequently, this results in more informative PRS for individuals with EUR ancestry but biased PRS for individuals from other ancestries.

Abigail Loe

First-year Master's Student
Biostatistics

Just Statistics: In the Dark, Statistical Analysis, and the Failure of the Justice System

3

On July 16th, 1996, someone walked into Tardy Furniture in Winona, Mississippi, and killed four employees. During an investigation which stressed and shocked the quiet community, District Attorney Doug Evans eventually accused, arrested and tried Curtis Flowers, a 26-year-old Black man. Curtis was first tried in 1997, found guilty, and sentenced to death in Mississippi's Fifth Circuit Court District. He won an appeal based on prosecutorial misconduct, and since then Doug Evans retried Curtis five times for the same crime. During Curtis Flowers' many trials, his defense team alleged racial bias in the seating of a jury. They were frequently unable to prove it in the trial court, but often won on appeal. Statistics however can show that something more than just chance has dictated the actions of a prosecutor, or the seating of a jury. Curtis Flowers' case shows that data has the potential to reveal the bias usually inherent in U.S. society, but a consistent disregard for statistical methods by the U.S. legal system has sidelined an avenue for rigorous proof. In this presentation, I examine the role that statistics can play in the law, and the ways in which the law has sidelined mathematical methods.

Anandkumar Patel

Second-year Master's Student
Statistics & D3 Center

Collaborator
Daniel Almirall

Standardized Effect Sizes for the Comparison of the Embedded, Clustered Adaptive Interventions in Clustered SMARTs

4

In many fields, such as in medicine and education, it is often necessary to make decisions about how best to intervene sequentially at the cluster level (e.g., at the level of a clinic or classroom), in a way that adapts and re-adapts the intervention over time, depending on the evolving needs of the cluster over time (including the cluster's response to prior intervention). Clustered Adaptive Interventions (CAIs) provide clinicians, educators, or other policymakers a guide to making such sequential, clustered intervention decisions. Often, however, there are open scientific questions preventing scientists from recommending a particular CAI. Clustered, sequential multiple assignments randomized trials (Clustered SMARTs) are one type of experimental design that can be used to answer such questions to develop highly effective CAIs. In SMARTs, randomization occurs at multiple stages corresponding to critical intervention decision points. Each randomization allows researchers to investigate and learn how to best adapt (potentially re-adapt) the intervention strategy at the cluster level while measuring the outcome at the individual level. Typically, Clustered SMARTs have a number of embedded CAIs, by design. A common primary aim in a SMART is the comparison of these embedded CAIs. This manuscript contributes to the statistical literature on clustered SMARTs by (i) defining effect sizes for the comparison of embedded CAIs in clustered SMARTs, deriving methods for (ii) estimating the effect size, and (iii) constructing confidence intervals for the estimated effect size. The methods are illustrated using data from a study that seeks to understand how best to implement Cognitive Behavioral therapy in high schools in Michigan.

Donald Scott

Junior
Statistics

Collaborator
Daniel Almirall

Discussion and Implementation of the Intra-Cluster Correlation in the Design and Analysis of Clustered SMARTs

5

In many fields, such as in medicine and education, it is often necessary to make decisions about how best to intervene sequentially at the cluster level (e.g., at the level of a clinic or classroom), in a way that adapts and re-adapts the intervention over time, depending on the evolving needs of the cluster over time (including the cluster's response to prior intervention). Clustered Adaptive Interventions (CAIs) provide a guide to making such sequential, clustered intervention decisions. Clustered, sequential multiple assignments randomized trials (Clustered SMARTs) are one type of experimental design that can be used to answer questions regarding the development of highly effective CAIs. In SMARTs, randomization occurs at multiple stages corresponding to critical intervention decision points. Each randomization allows researchers to investigate and learn how to best adapt (potentially re-adapt) the intervention strategy at the cluster level, while measuring the outcome at the individual level. The ICC (Inter Cluster Correlation) plays a crucial role in the design and analysis of Clustered Smarts. The ICC provides information about the amount of variance within the study related to the grouping of clusters. This manuscript contributes to the statistical literature on clustered SMARTs by (i) describing a method of calculating the ICC from existing data from a Clustered SMART and (ii) methods of calculating the ICC using existing data to inform the design of a clustered SMART.

Fatema Shafie Khorassani

Third-year PhD Student
Biostatistics

Collaborators

Jeremy M.G. Taylor

Xu Shi

Data Fusion for Time-to-Event Outcomes

Despite significant reductions in cancer mortality over the past three decades, racial disparities in cancer-specific mortality persist. Studying factors associated with these observed disparities requires data on many variables, including demographics, healthcare access, socioeconomic status, and comorbidities. There are existing national cancer surveillance databases that each collect parts of the information needed for studying racial disparities in cancer. Integrating data from multiple sources allows us to study associations between race and cancer-specific mortality over time adjusted for important confounders. Existing data integration methods do not consider time-to-event outcomes, hence are not applicable to studying cancer-specific mortality. Data integration methods for time-to-event outcomes can have many applications, including improving risk predictions, adjusting for dependent censoring, finding new associations, adjusting for unmeasured confounding, and improving the efficiency of analyses.

We propose a doubly robust regression method for data fusion with a time to event outcome. Data fusion is a particularly challenging problem in data integration, in which no subject has complete data on all the covariates and outcome. Some existing missing data methods have been extended to the setting of data fusion; however, they do not account for time-to-event outcomes. We present a method for regressing a time-to-event outcome on a set of covariates from two integrated datasets that include some overlapping variables. We will present a class of doubly robust estimators which are unbiased if either the data source model or the model of the unobserved covariates is specified correctly. Through simulation studies we will present the bias and coverage of our estimators under correctly specified and misspecified models and will apply the method to integrate cancer-specific mortality information from the Surveillance, Epidemiology, and End Results (SEER) Program with confounders collected in the National Cancer Database (NCDB) that are not available in SEER.

Jiahao Shi

First-year PhD Student
IOE

Collaborators

Albert S. Berahas

Zihong Yi

Baoyu Zhou

Accelerating Stochastic Sequential Quadratic Programming for Equality Constrained Stochastic Optimization using Predictive Variance Reduction

We propose a variance reduction method for solving equality constrained stochastic optimization problems. Specifically, we develop a method based on the sequential quadratic programming paradigm that utilizes variance reduction in the gradient approximation via stochastic variance reduction gradient (SVRG). We prove exact convergence in expectation to first order stationarities with non-diminishing stepsize sequences. Finally, we demonstrate the practical performance of our proposed algorithm of standard constrained machine learning problems.

Xianlin Sun

Second-year Master's Student
Statistics

Collaborators

Shasha Zou

Yang Chen

Hu Sun

Jiaen Ren

Machine Learning Forecast and Statistical Exploration of Equatorial Ionization Anomaly Based on Total Electron Content

6

The ionosphere total electronic content (TEC), derived from multi-frequency Global Navigation Satellite System (GNSS) receiver, has been one of the most popular datasets in ionosphere research academia. The new advances in the completion of TEC maps and the forecast of TEC data by the modern ML(Machine Learning) algorithms have significantly leveraged its usability. While observing the TEC data, significant equatorial ionization anomaly (EIA) phenomenon displays that observably high TEC values occur around the magnetic equator lasting for a noticeable time, showing two strips each on one side of the equator. As we marched into multi-GNSS era, a new frontier of combining the traditional space science and the cutting-edge statistical learning to make a breakthrough in the specification and forecasting of EIA phenomenon has emerged. In this project, we aim at specifying EIA phenomena by automatically identifying its location and statistically describing its properties. We adopt Gaussian Mixture Model (GMM) with relatively free number of peaks to specify the EIA phenomenon and a series of state-of-the-art ML algorithms will be used to forecast local, regional and global EIA behavior. We automatically identify the EIA peaks, evaluating the peak TECs and prominences, and other key features, such as peak to equator distances and hemispheric asymmetry. Based on these EIA properties obtained, we could further explore the evolution of EIA peaks and the frequency, duration, intensity and periodicity of EIA bifurcation by constructing a state-of-the-art ML model based on the constructed EIA database as well as data indicating space weather conditions, for instance, solar wind and FISM solar radiation measurements, etc.

Sahita Manda

Junior
Psychology

Collaborators

Elizabeth Buvinger

Shichi Dhar

Harika Veldanda

Experience of stigma and its relationship to identification with the neurodiversity model for Indian parents of children with autism spectrum disorder

7

It is widely recognized that individuals with autism spectrum disorder (ASD) and their families continue to face extensive stigma and that much of the current research on ASD is deficit-focused. Diversity and inclusion perspectives are emerging, but there is less of a focus on how stigma affects the adoption of these approaches. In collaboration with the University of Michigan Department of Psychology and the national Indian organization Action For Autism, our research aims to understand the experience of stigma and its relationship to identification with the neurodiversity model for Indian parents of children with ASD. The study was carried out by administering online surveys through the platform of Qualtrics to Indian parents residing in India (N=56). This study explores the extent to which Asian value adherence, child functioning, and perceived ASD stigma contribute to parental alignment with the neurodiversity model. It also investigates the ways in which alignment with the model affects parental stress, isolation from family and friends, parenting goals, identification of child's strengths, and positive perceptions about raising a child with ASD. Preliminary findings demonstrate statistically significant correlations between a child's ASD behaviors, perceived ASD stigma, parental stress, and isolation from family and friends. A more complex mediation model of the effects of neurodiversity alignment on these variables will be presented and will have implications for the adoption of strength-based practices and the reduction of stigma associated with ASD within different cultural contexts.

Ziping Xu

Fourth-year PhD Student
Statistics

Collaborator

Ambuj Tewari

On the Statistical Benefits of Curriculum Learning

Curriculum learning (CL) is a commonly used machine learning training strategy. However, we still lack a clear theoretical understanding of CL's benefits. In this paper, we study the benefits of CL in the multitask linear regression problem under both structured and unstructured settings. For both settings, we derive the minimax rates for CL with the oracle that provides the optimal curriculum and without the oracle, where the agent has to adaptively learn a good curriculum. Our results reveal that adaptive learning can be fundamentally harder than the oracle learning in the unstructured setting, but it merely introduces a small extra term in the structured setting. To connect theory with practice, we provide justification for a popular empirical method that selects tasks with highest local prediction gain by comparing its guarantees with the minimax rates mentioned above.

Guanghao Zhang

First-year PhD Student
Biostatistics

Collaborators

Xiaoou Li

Tianxi Cai

Xu Shi

A bipartite graph model for medical code mapping between healthcare systems

It is notorious that electronic health records (EHR) data do not talk to each other. Due to financial incentives and differential care practice, the same clinical concept can often be described by alternative medical codes in different healthcare systems, leading to idiosyncratic "dialects" of EHRs across systems. Variability in medical coding has been observed for decades and can degrade model performance when models are applied to a new healthcare system. To facilitate data integration and improve model transportability, we adopt principles in how humans talk to each other and develop a data-driven method that automatically maps medical codes between two systems. We formulate a bipartite graph model that, unlike existing language translation methods, is naturally symmetric, accommodates all patterns such as one-to-one and one-to-many mappings, and does not require prior knowledge on code mapping or grouping. We demonstrate the validity of our proposed medical code mapping method through a simulation study and an application study of mapping ICD codes between two healthcare systems.

Ruixuan Zhang

-

Civil and Environmental Engineering

Collaborators

Sara Masoud

Neda Masoud

Is the Car Following Behaviour of Human Drivers Affected when Following Autonomous Vehicles?

8

In this work, we use naturalistic driving data from NGSIM and Lyft Level 5 prediction datasets to evaluate the potential effects of autonomous vehicles (AVs) on human drivers' car-following behavior. We use time headway time series as a proxy to capture the car following behaviors of human drivers. A nested fixed model is developed to find possible changes when human drivers are following different types of vehicles (human-driven vehicles or AVs). The factors included in this model are the platoon structure (Legacy-Following-Legacy and Legacy-Following-Autonomous-Vehicle), road type (freeway and urban), time period (morning and afternoon), and lane (right, middle, and left). Results indicate a statistically significant difference between the car following behavior of drivers when they follow a human-driven vehicle, compared to an AV. This change in the car following the behavior of drivers has manifested in the form of a reduction in the mean and variance of time headways when human drivers follow an AV. These findings can bridge the gap between anticipated and real-world impacts of AVs on traffic streams as well as roadway safety and capacity.

Yongwen Zhuang

Fourth-year PhD Student Biostatistics

Collaborators

Bhramar Mukherjee

Seunggeun Lee

A matrix completion approach for potential disease risk prediction

Identification of at risk population is important for early-stage disease prevention. While increasing number of large-scale GWAS studies help support the risk prediction of various diseases through polygenic risk scores (PRS), the rapidly increasing PHEWAS studies provides further insight into the prediction of rare diseases with the use of multiple PRS across the phenome spectrum. However, two major challenges remain in utilizing phenome-wide information for risk prediction. Firstly, solutions remain unclear regarding the "unsupervised" scenario where no or little phenotypic information is available for the model calibration step of existing methods. Secondly, the existing cross-phenotype risk prediction methods are often trained in a disease-by-disease fashion, leading to increased computation burden when a larger number of diseases were of interest. We propose a computationally efficient matrix completion approach to identify potential at-risk individuals for diseases with small amount of case information by combining prior knowledge about individual similarity (constructed using genetic relatedness and health related features) and disease similarity available in various external data sources. Through simulations and analysis of biobank data, we show that the proposed method outperforms benchmark methods in terms of prediction accuracy and AUC.

POSTER SESSION

March 10th, 2:30pm – 4:00pm

Hussey

See also abstract 1 to 8 of the Speed Session

Prayag Chatha

Second-year PhD Student
Statistics

Collaborators

Jessica Mellinger

Jeffrey Regier

Early Detection of Alcoholic Liver Disease in the Optum Claims Dataset with Transformers

9

Alcoholic liver disease (ALD) is a leading cause of liver-related death world-wide. Unfortunately, ALD is often diagnosed too late for effective intervention. The Optum Claims dataset contains billing codes for the employee-sponsored insurance claims of millions of individuals—a vast amount of observational data about a general population, including patients with ALD. As a patient inter-acts with the medical system over time, they generate a detailed sequence of ICD diagnostic codes, lab results, and drug prescriptions in Optum Claims. A transformer is a deep learning architecture that excels at modeling long-range sequential dependencies in sequential data through self-attention. Unlike a re-current neural network, a transformer admits parallel computation for efficient training. We developed a transformer-based model called "tf-md" that differentiates early-stage and late-stage ALD based on Optum Claims data. tf-md achieved a validation AUROC of 0.689, whereas a fully-connected "bag of words" neural network model had a best AUROC of 0.674. The latter model has access only to the frequency of codes, not their sequence position, suggesting that the ordering of a patient's insurance claims contains information that helps to detect ALD early.

Irena Chen

Fourth-year PhD Student
Biostatistics

Collaborators

Zhenke Wu

Siobán D. Harlow

C. A. Karvonen-Gutierrez

Michelle M. Hood

Michael R. Elliott

Modeling Individual Variability to Predict Health Outcomes: A Joint Hierarchical Bayesian Approach

10

Longitudinal biomarker data and cross-sectional outcomes are routinely collected in modern epidemiology studies, often with the goal of informing tailored early intervention decisions. For example, hormones such as estradiol (E2) and follicle-stimulating hormone (FSH) may predict changes in womens' health during the midlife. Most existing methods focus on constructing predictors from mean marker trajectories. However, subject-level biomarker variability may also provide critical information about disease risks and health outcomes. In this paper, we develop a joint model that estimates subject-level means and variances of longitudinal predictors to predict a cross-sectional health outcome. Simulations demonstrate excellent recovery of true model parameters. The proposed method provides less biased and more efficient estimates, relative to alternative approaches that either ignore subject-level differences in variances or perform two-stage estimation where estimated marker variances are treated as observed. Analyses of women's health data reveal that larger variability of E2 and higher mean levels of E2 and FSH are associated with higher levels of fat mass change across the menopausal transition.

Seokhyun Chung

Fourth-year PhD Student
IOE

Collaborator

Raed Al Kontar

Federated Condition Monitoring Signal Prediction with Improved Generalization

11

Revolutionary advances in Internet of Things technologies have paved the way for a significant increase in computational resources at edge devices that collect condition monitoring (CM) data. This poses a significant opportunity for federated analytics which exploits edge computing resources to distribute model learning, reduce communication traffic and circumvent the need to share raw data. In this paper we study CM signal prediction where operating units, that have data storage and computational capabilities, jointly learn models without sharing their collected CM signals. Specifically, we first propose a framework for CM signal prediction and introduce a federated approach that tries to improve generalization by encouraging flat solutions through distributed computations. Then, a personalization approach is proposed to adapt the learned model to new clients without losing old knowledge. We examine our proposed framework on CM signals from aircraft turbofan engines under three realistic federated CM scenarios. Experimental results highlight the advantageous features of the proposed approach in improving generalization while decentralizing model inference.

Dylan Glover

Second-year Master's Student
Statistics

Collaborator

Daniel Iong

Forecasting Geomagnetically Induced Currents at the Ottawa Magnetometer Station using ACE Variables

12

Geomagnetic storms occur when high-speed solar wind induces fluctuations in the Earth's geomagnetic field. We consider a forecasting model of these fluctuations as a proxy for geomagnetically induced currents (GICs) produced during storms, which can damage ground-based infrastructure and result in regionalized power and utility blackouts. The target quantity in this study was the maximum of the horizontal component of the geomagnetic field dB/dt over a 20-minute interval at the Ottawa SuperMAG magnetometer station. We chose to study storm times only so that the model learns storm behavior, rather than the characteristics of idle time periods. The 1-min resolution interplanetary magnetic field and plasma data were gathered by NASA's Advanced Composition Explorer (ACE), and then smoothed, to forecast the target with one hour lead time. The random forest model was chosen to enable post-hoc interpretability using SHAP values, which explain the contribution of individual features to each test set observation's point estimate prediction. Preliminary results indicate reasonable root mean square error of approximately 15 nT (nanotesla) on the test set, so this project will focus on the model as a starting point for quantifying the performance of various modeling choices under various data processing and splitting regimes. In future work, this model will also be compared to models trained on data originating from other stations at various latitudes, to determine how location may influence the features driving held-out test set performance.

Yiling Huang

Senior
Mathematics & Statistics

Collaborator

Mark Fredrickson

Balance Assessment of Matched Data with Multiple Treatment Levels

13

Identifying and estimating causal effects of treatments is of significant research interest. In doing so, similar data are oftentimes matched into one stratum, and subsequent inferences of causality are carried out based on these strata. In particular, when the data are from observational studies, properly matching observations by their treatment assignment probabilities are especially important for removing potential selection bias induced by selecting observations that receive specific treatments in a non-randomized fashion. Therefore, it is an important task to evaluate whether matching was done properly, that is, whether the covariates are equally distributed in different treatment groups given the matching information. Traditional methods of matching evaluation involve visually investigating summary statistics, such as the standardized mean difference, by covariate, but lack uncertainty quantification of the conclusion and are less convenient compared to an omnibus test that checks matching validity for all covariates one-shot. We propose a hypothesis test that expresses treatment assignment probabilities by an adjacent category logistic regression model and provides an omnibus test of matching for all covariates by testing the global null $\beta = 0$ in the language of regression models. In this thesis, we adopt a $\chi 2$ approximation of the asymptotic distribution of the test statistic, inspired by the Rao score test. An application of the test indicates the matching results produced by a matching algorithm can be further improved.

Roman Kouznetsov

Third-year PhD Student
Statistics

Collaborators

Jackson Loper

Jeffrey Regier

deepST: A Graph Convolutional Autoencoder for Spatial Transcriptomics

14

Spatial transcriptomics (ST) measures gene expression for individual cells and pairs these measurements with the positions of cells within a tissue sample. This opens the door for statistical methods to explore how neighboring cells interact. The statistical structure of these interactions can be investigated by posing prediction problems. For example, we can see which subsets of genes in neighboring cells are most predictive of gene expression in target cells. We can infer conditional independence structures by comparing prediction accuracy obtained from different subsets. Existing methods pursuing this vision use fixed-dimensional summaries of the attributes of neighboring cells, ignoring the number of neighbors and the interactions among them. We here propose deepST, a denoising graph convolutional autoencoder that accounts for these subtleties. For a large MERFISH hypothalamus dataset, deepST imputes missing expression levels for response genes more accurately than other state-of-the-art methods including gradient boosting, attaining a 8.7% reduction in absolute error. We also find that gradient boosting itself outperforms existing methods in this domain such as "Mixture of Experts for Spatial Signaling genes Identification'", attaining a 7.2% reduction in absolute error.

Subha Maity

Fourth-year PhD Student
Statistics

Collaborators

Debarghya Mukherjee

Mikhail Yurochkin

Yuekai Sun

Does enforcing fairness mitigate biases caused by subpopulation shift?

15

Many instances of algorithmic bias are caused by subpopulation shifts. For example, ML models often perform worse on demographic groups that are underrepresented in the training data. In this paper, we study whether enforcing algorithmic fairness during training improves the performance of the trained model in the target domain. On one hand, we conceive scenarios in which enforcing fairness does not improve performance in the target domain. In fact, it may even harm performance. On the other hand, we derive necessary and sufficient conditions under which enforcing algorithmic fairness leads to the Bayes model in the target domain. We also illustrate the practical implications of our theoretical results in simulations and on real data.

Robert Malinas

Fourth-year PhD Student
EECS

Collaborators

Benjamin D. Robinson

Alfred O. Hero III

Detecting Changes in the Covariance Structure of a High-Dimensional Random Process

16

Stationarity is a property often assumed of random samples in a variety of statistical techniques. Our purpose is to determine whether an independent random sample is stationary up to second order, i.e., whether the covariance matrix of the observations is homogeneous throughout the sample. Given a sample of size T of an N-dimensional random process, where T is on the order of N, we consider the presence of a single change point $2 \le r \le T$ such that the observations indexed $\{1, ..., r\}$ are i.i.d. and have a different covariance matrix than the observations indexed $\{r + 1, ..., T\}$, also assumed to be i.i.d. Our procedure is to first determine whether $r = T$ and, if not, estimate r. Using random matrix theory and free probability theory, we develop a statistic S(t) such that S(t) converges to 0 in probability, in the proportional growth limit of random matrix theory with respect to N and T, for all $2 \le t \le T$ if $r = T$. Furthermore, if $r < T$, we show that S(r) is asymptotically greater than or equal to S(t) with high probability for every $2 \le t \le T$. This yields a procedure by which we can detect and estimate the change point by thresholding. Due to the universality of the random matrix theory used, these results are shown under mild regularity conditions; in particular, we assume only the existence of second moments. Finally, we discuss convergence rates in probability and the statistical performance of the associated test.

Stephen Salerno

Fourth-year PhD Student
Biostatistics

Collaborator

Yi Li

A New Deep Learning Approach for Predicting Survival Processes in the Presence of Semi-Competing Risks

17

Many survival processes involve a non-terminal (e.g., disease progression) and a terminal (e.g., death) event, which form a semi-competing risk relationship, i.e., the occurrence of the non-terminal event is subject to the terminal event. Deep learning has emerged as a powerful tool for survival prediction; however, limited work has been done to predict multi-state or competing risk outcomes, let alone semi-competing outcomes. We propose a new deep learning framework for predicting semi-competing risk outcomes based on the illness-death model, a compartment-type model for transitions between states, which allows us to estimate patient-specific transition hazards, including the sojourn time between events, and patient frailty. As deep learning can recover non-linear risk scores, we test our method predicting simulated risk surfaces of varying complexity. We apply our method to the Boston Lung Cancer Study, where we study the impact of clinical and genetic predictors on disease progression and mortality, and the Michigan Medicine Precision Health initiative, where we quantify risks for COVID-19 hospitalization and mortality.

Zeyu Sun

Third-year PhD Student
EECS

Collaborators

Monica Bobra

Yu Wang

Hu Sun

Yang Chen

Alfred Hero

Predicting Solar Flares Using CNN and LSTM on Two Solar Cycles of Active Region Data

18

We consider the flare prediction problem that distinguishes flare-imminent active regions that produce an M- or X-class flare in the future 24 hours, from quiet active regions that do not produce any flare within ±24 hours. Using line-of-sight magnetograms and parameters of active regions in two data products covering Solar Cycle 23 and 24, we train and evaluate two deep learning algorithms—CNN and LSTM—and their stacking ensembles. The decisions of CNN are explained using visual attribution methods. We have the following three main findings.
(1) LSTM trained on data from two solar cycles achieves significantly higher True Skill Scores (TSS) than that trained on data from a single solar cycle with a confidence level of at least 0.95.
(2) On data from Solar Cycle 23, a stacking ensemble that combines predictions from LSTM and CNN using the TSS criterion achieves significantly higher TSS than the "select-best" strategy with a confidence level of at least 0.95.
(3) A visual attribution method called Integrated Gradients is able to attribute the CNN's predictions of flares to the emerging magnetic flux in the active region. It also reveals a limitation of CNN as a flare prediction method using line-of-sight magnetograms: it treats the polarity artifact of line-of-sight magnetograms as positive evidence of flares.

Leyao Zhang

First-year PhD Student
Biostatistics

Collaborators

Wen Wang

Mengtong Hu

Alan P. Baptist

Peter X.K. Song

Adaptive learning of relevant questions from a questionnaire via best subset algorithms

19

Questionnaire is one of the oldest and most widely used instruments in practice to measure variables relevant to certain traits of interest that cannot be easily measured by physical devices. This paper is bonded with a cohort study of elderly asthma patients in that we aim to examine associations between clinical outcomes and quality of life (QoL). In many practical studies, including our asthma clinical study, the scope of a questionnaire (e.g. QoL) is unfit to a new study population that appears different from the original population used for either questionnaire development or validation. As a result, items in a questionnaire may or may not be of relevance to the new study population. In our analysis, we consider a supervised learning method to identify a subset of questions whose summary score is maximally associated with a specific clinical outcome under investigation. The resultant set of selected items gives an optimal summary metric of the questionnaire, which improves both statistical power and clinical interpretation. Our item extraction procedure is built upon the best subset algorithm implemented by a mixed integer programming, which enjoys both theoretical guarantee of selection consistency and flexibility of handling non-responses. This best subset algorithm is first evaluated by extensive simulation studies with comparisons to existing methods, and then applied to derive tailored QoL scores adaptive to two clinical outcomes of lung function measure (FEV1) and asthma control test (ACT), respectively, among elderly people with persistent asthma.

ORGANIZING COMMITTEE

Lap Sum Chan

PhD Student
Biostatistics

Curtiss Engstrom

PhD Student
Program in Survey and
Data Science

Simon Fontaine

PhD Student
Statistics

Cheoljoon Jeong

PhD Student
Industrial and Operations
Engineering

Alexander Ritchie

PhD Student
Electrical and Computer
Engineering

THANKS

The *Organizing Commitee* would like to thank all the people and institutions that allow MSSISS 2022 to be the great event that it is.

First, we would like to thank our faculty mentors, Raed Al Kontar (IOE), Johann Gagnon-Bartsch (Statistics), Timothy Johnson (Biostatistics), Brady West (Survey Methodology) and Clayton Scott (EECS) for their precious advice and their contribution to judging and awarding prizes.

Second, we would like to thank the 2021 MSSISS organizers, Seokhyun Chung (IOE), Yijun Li (Biostatistics), Zeyu Sun (EECS), Ziping Xu (Statistics) and Xinyu Zhang (Survey Methodology), as well as the 2021 faculty mentor Ed Ionides (Statistics), who helped with the transition and provided useful insight and material for the forthcoming adventure.

Third, we would like to thank Judy McDonald, from the Statistics department, for all the administrative support, Lindsay Sorgenfrei, from Michigan League, for all the logistical help and Holly McCamant, from the department of Biostatistics, for managing the finances, all of which were essential to ensure everything runs smoothly.

Fourth, we would like to thank Renee Li (EECS, MDST) who helped design and organize the Data Challenge that unfortunately did not attract enough participants to take place.

Fifth, we would like to thank all our sponsors, listed on the back cover, without which none of this would be possible. They allow us to keep registration free and still offer a unique and welcoming experience to all.

Lastly, we would like to thank our speakers and presenters for agreeing to be part of this amazing experience.

MSSISS 2022

MARCH 10–11

MICHIGAN LEAGUE 2nd FLOOR

911 N University Ave
Ann Arbor, MI, USA
48109

SPONSORS

LSA STATISTICS
UNIVERSITY OF MICHIGAN

RACKHAM GRADUATE SCHOOL
UNIVERSITY OF MICHIGAN
Dean's Strategic Initiatives Fund

INSTITUTE FOR SOCIAL RESEARCH
SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

BIOSTATISTICS

IOE INDUSTRIAL & OPERATIONS ENGINEERING
UNIVERSITY OF MICHIGAN

ECE ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY OF MICHIGAN
With additional contributions from Professors
Laura Balzano, Alfred Hero and Clayton Scott

INSTITUTE FOR SOCIAL RESEARCH
PROGRAM IN SURVEY AND DATA SCIENCE
UNIVERSITY OF MICHIGAN

LSA COLLEGE OF LITERATURE, SCIENCE, AND THE ARTS
UNIVERSITY OF MICHIGAN

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

CONTACT

https://sites.lsa.umich.edu/mssiss/
mssiss2022-contact@umich.edu