Research Alumni Symposium 2024

October 31 - Nov 2, 2024

Titles and abstracts of oral presentations

Last updates October 29, 2024

Omar Aguilar
Charles Schwab

Title. Industry Panel
Abstract. In this industry panel we will have a conversation with leading experts to discuss the transformative applications of statistical science and Bayesian Statistics across diverse disciplines, including finance, public policy, and bioinformatics. As data-driven decision-making becomes increasingly vital, the panel will delve into innovative statistical tools and methodologies that enhance analysis and forecasting in these fields. Panelists will share insights on their current research and share their experience addressing sophisticated industry problems. Attendees will gain a deeper understanding of the interdisciplinary nature of Bayesian Statistics and its critical role in addressing complex challenges in todays data-rich environment. Join us for an engaging discussion that highlights the power of statistical tools in driving impactful solutions across various sectors.

Lindsay Berry
Berry Consultants

Title. Bayesian modeling in COVID-19 clinical trials
Abstract. The coronavirus pandemic created an urgent need for effective treatments for severely ill patients at a time of significant uncertainty. Adaptive platform trials, rapidly implemented during the outbreak, played a crucial role in identifying both life-saving therapies and those with limited efficacy. In this talk, Lindsay will share insights from her transition from Duke to working as a statistician on Bayesian clinical trials, which randomized thousands of hospitalized COVID-19 patients and assessed numerous potential treatments. The focus of the talk will be on the key role of Bayesian modeling in these trials in addressing the urgency and uncertainty of the pandemic.

Fan Bu
University of Michigan

Title. Bayesian safety surveillance on federated observational health data
Abstract. Post-market safety surveillance is an integral part of mass vaccination programs. Typically relying on sequential analysis of real-world health data as they accrue, safety surveillance is challenging due to sequential multiple testing and biases induced by residual confounding. The current standard approach based on the maximized sequential probability ratio test (MaxSPRT) fails to satisfactorily address these practical challenges and it remains a rigid framework that requires a pre-specified surveillance schedule. We develop an alternative Bayesian surveillance procedure that addresses both challenges using a more flexible framework. Our approach sequentially estimates vaccination effect on adverse events of interest and corrects estimation bias by simultaneously analyzing a large set of negative control outcomes through a Bayesian hierarchical model. To detect safety signals, we use posterior probabilities of hypotheses computed via Markov chain Monte Carlo sampling. Through an empirical evaluation using six US observational healthcare databases covering more than 360 million patients, we benchmark the proposed procedure against MaxSPRT on testing errors and estimation accuracy, under two epidemiological designs, the historical comparator and the self-controlled case series. We demonstrate that our procedure substantially reduces Type 1 error rates, maintains high statistical power, delivers fast signal detection, and provides considerably more accurate estimation. As an effort to promote open science, we present all empirical results in an R ShinyApp and provide full implementation of our method in the R package EvidenceSynthesis.

Carlos Carvalho
University of Texas, Austin

Title. The Book that Refuses to Die!
Abstract. For 20+ years I have used and found inspiration on ”Bayesian Forecasting and Dynamic Models”. In this talk I will discuss how ideas found in Mike's book have permeated my research career to this day and are still incredibly relevant in the world.

Lorin Crawford
Microsoft Research

Title. Statistical challenges and opportunities in defining, modeling, and targeting cell state in cancer
Abstract. Project Ex Vivo is a joint cancer research collaboration between Microsoft and the Broad Institute of MIT and Harvard. Our group views cancers as complex (eco)systems, beyond just mutational variation, that necessitate systems-level understanding and intervention. In this talk, I will discuss a series of multimodal statistical and deep learning approaches to understand accurate representations of tumors by integrating genetic markers, expression state, and microenvironmental interactions. These representations help us precisely define and quantify the trajectory of each tumor in each patient. Our ultimate objective is to more effectively model cancer ex vivo outside the body in a patient-specific manner. In doing so, we aim to unlock the ability to better stratify patient populations and identify therapies that target diverse aspects of human cancers.

Marco Ferreira
Virginia Polytechnic Institute and State University

Title. Bayesian genome-wide iterative fine-mapping
Abstract. Fine-mapping seeks to identify causal variants in genomic regions of interest previously identified by genome-wide association studies (GWAS). But because fine-mapping is performed separately from GWAS, fine-mapping does not extract as much information as possible from the data. From a statistical point of view, this is a $p \gg n$ variable selection problem for (generalized) linear mixed models where there are $O(10^5)$ regressors and $O(10^3)$ observations. Here we present novel Bayesian Genome-wide Iterative fiNe-mApping methods for Gaussian data (GINA) and non-Gaussian data (GINA-X). GINA and GINA-X efficiently extract information from GWAS data by iterating two steps: a screening step and a model selection step. The screening step provides a list of candidate genetic variants and an estimate of the proportion of null genetic variants. After that, the model selection step searches the model space defined by the list of candidate genetic variants and uses the estimated proportion of null genetic variants to appropriately control for genome-wide multiplicity. A simulation study shows that, when compared to competing fine-mapping methods, GINA and GINA-X reduce false discovery rate and increase recall of true causal genetic variants. We illustrate the usefulness and flexibility of GINA and GINA-X with case studies on alcohol use disorder and breast cancer.

Mengyang Gu
University of California, Santa Barbara

Title. Fast ab initio uncertainty quantification and data inversion for dynamical systems
Abstract. Estimating parameters from data is a fundamental problem, which is customarily done by minimizing a loss function between a model and observed statistics. In this talk, we discuss another paradigm termed the ab initio uncertainty quantification method, for improving loss-minimization estimation in two steps. In step one, we define a probabilistic generative model from the beginning of data processing and show the equivalence between loss-minimization estimation and a statistical estimator. In step two, we develop better models, more efficient estimators, or faster algorithms to improve the estimation. To illustrate, we introduce two approaches to estimate dynamical systems, one in Fourier analysis of microscopy videos, and the other in inversely estimating the particle interaction kernel from trajectory. In the first approach, we show that differential dynamic microscopy, a scattering-based analysis tool that extracts dynamical information from microscopy videos, is equivalent to fitting temporal auto-covariance in the spatial Fourier domain, based on a latent factor model we constructed. We derive likelihood-based inference and reduce the computational complexity to pseudolinear order of the number of observations by utilizing the generalized Schur algorithm for Toeplitz covariances. In the second approach, we develop a new method called the inverse Kalman filter which enables accurate matrix-vector multiplication between a covariance matrix from a dynamic linear model and any real-valued vector with a linear computational cost. These new approaches outline a wide range of applications, such as probing optically dense systems, automated determination of gelation time, and estimating cellular interaction for fibroblasts on liquid crystalline substrates.

Jingchen Monika Hu
Vassar College

Title. Mechanisms for Global Differential Privacy under Bayesian Data Synthesis
Abstract. We review, propose, and compare several Bayesian data synthesizers with different differential privacy guarantees that can be used by data stewards for microdata dissemination with privacy protection. The pseudo posterior mechanism achieves an asymptotic differential privacy guarantee and a variant of it can provide faster convergence. The newly proposed censoring mechanism embedded in the pseudo posterior mechanism censors the pseudo likelihood of every record within $[e^{-\epsilon/2}, e^{\epsilon/2}]$, which provides a stronger, non-asymptotic differential

privacy guarantee. Through a series of simulation studies with bounded, univariate data and an application to sample of the Survey of Doctoral Recipients where a beta regression synthesizer is utilized, we demonstrate that the pseudo posterior mechanism creates synthetic data with the highest utility at the price of a weaker, asymptotic privacy guarantee, while the censoring mechanism embedded in the pseudo posterior mechanism produces synthetic data with a stronger, non-asymptotic privacy guarantee at the cost of slightly reduced utility. The perturbed histogram is included for comparison.

Gabriel Huerta
Sandia National Laboratories

Title. Bayesian examples at Sandia National Laboratories
Abstract. In this talk, I'll introduce two problems at Sandia where we had leveraged Bayesian methods and ideas. I'll demonstrate how we had been performing Bayesian model calibration for experiments executed at Sandia's Z-machine, a power pulsed facility in New Mexico that uses high magnetic fields associated with high electrical currents to produce high temperatures, high pressures and conditions found nowhere else on Earth. The second example connected me back to West and Harrison (1997) and related work on DLMs. Ill show how a multivariate space time dynamic model (MV-STDLM) has been applied to characterize the atmospheric impacts following the 1991 Mt. Pinatubo eruption which resulted in a massive increase of sulfate aerosols in the atmosphere, absorbing radiation and leading to global changes in surface and stratosphere temperatures.

Didong Li
University of North Carolina at Chapel Hill

Title. Spatial Transcriptomics: Opportunities and Challenges
Abstract. In this talk, I will provide a brief overview of the emerging field of Spatial Transcriptomics (ST), which integrates spatial information with gene expression data to offer new insights into tissue organization and function. I will introduce our recently curated dataset, STimage-1K4M, which pairs high-resolution histopathology images with gene expression data, offering new opportunities for statistical analysis. The talk will also highlight key challenges such as high dimensionality, batch effects, heterogeneity, and multi-modal integration, and explore exciting directions for future work in developing models and tools that can leverage the full potential of ST data.

Lizhen Lin
The University of Maryland

Title. Statistical foundations of deep neural network models
Abstract. As deep learning has achieved breakthrough performance in many modern tasks, significant efforts have been made to understand thetheoretical foundations of such models.This talkwillfocus on providing theoretical underpinning for deep neural network (DNN) models through the lens of statistical theory. From a statistical viewpoint, a deep learning model can be largely considered as a nonparametric function or distribution estimation where the underlying function or distribution is parametrized by a deep neural network. One key question is understanding why DNN models perform so well in high-dimensional settings and seem to be able to circumvent the curse of dimensionalityan issue that classical nonparametric estimators typically face. In this talk, we will elucidate the statistical theory behind the superior performance of DNN-based estimators. The key takeaway, as informed by the theory, is that DNN models have the ability to adapt to the intrinsic structure of the data, whether in the context of deep supervised learning or deep unsupervised learning.

Hedibert Lopes
Insper - Institute of Education and Research

Title. Brazilian dynamic modellers
Abstract. This is not a technical talk, but I will revisit and connect some dots regarding how dynamic modeling became the bedrock of Bayesian in Rio de Janeiro. In order to do that, we go back to Harrison and Stevens (1976) and then West, Harrison and Migon (1985) and how Mike, Migon, Dani were responsible to creating a legion of dynamic modellers from Rio de Janeiro, including myself, Marco, Alexandra, Carlos, and many others after us.

Ken McAlinn
Temple University

Title. Is our view of Bayesian statistics not foundational enough?
Abstract. What does it mean to solve a problem as a Bayesian? Many approaches exist within the Bayesian sphere, but not all fully leverage the coherent framework the Bayesian view provides. What I learned from Mike is that asking what a (subjective) Bayesian "should" do, in terms of foundations, utilities, and decisions, is not only an esoteric philosophical exercise, but a fundamentally practical approach to solving problems as a Bayesian. I will discuss how Mike has influenced my re-

search, through my PhD thesis to current projects with my students.

Kelly Moran
Los Alamos National Laboratory

Title. Bayesian Modeling of Ion Temperature Diagnostics for Inertial Confinement Fusion
Abstract. Inertial Confinement Fusion (ICF) experiments aim to achieve controlled thermonuclear reactions by compressing fuel capsules with powerful lasers. Ion temperature (T) within the plasma drives reaction rates, energy yield, and overall fusion efficiency. Precise measurements of T allow researchers to tune the laser drive and target design to obtain higher yields. Current state of the art in ion temperature diagnostics rely on maximum likelihood based approaches to learn T. However, these methods can struggle to converge reliably, do not account for known information about the shape of T, and do not provide uncertainty about the estimated profile. We have developed a Gaussian Process (GP)-based Bayesian modeling framework that learns spatially varying profiles for temperature, along with other parameters, from 1D streak camera data. This approach has enabled the first physics observation of a plasma temperature profile from these data to date.

Long Nguyen
University of Michigan

Title. Dendrogram of mixing measures: Hierarchical clustering and model selection using finite mixture models
Abstract. We present a new method to summarize and select mixture models via the hierarchical clustering tree (dendrogram) constructed from an overfitted latent mixing measure. Our proposed method bridges agglomerative hierarchical clustering and mixture modeling. The dendrogram's construction is derived from the theory of convergence of the mixing measures, and as a result, we can both consistently select the true number of mixing components and obtain the pointwise optimal convergence rate for parameter estimation from the tree, even when the model parameters are only weakly identifiable. In theory, it explicates the choice of the optimal number of clusters in hierarchical clustering. In practice, the dendrogram reveals more information on the hierarchy of subpopulations compared to traditional ways of summarizing mixture models. Several simulation studies are carried out to support our theory. We also illustrate the methodology with an application to single-cell RNA sequence analysis. This work is joint with Dat Do, Linh Do, Scott McKinley and Jonathan Terhorst.

Jarad Niemi
Iowa State University

Title. Mike West Legacy through the Niemi Branch
Abstract. As an academic, our legacy is not only a product of our own direct work, but also the work of our academic descendants. According to the mathematics genealogy project, Mike has advised 42 PhD students and those students have advised another 62 PhD students. As Mike's advisee, he has shaped me into the Statistician I am today and can claim some credit (and no blame) for my work. In my 16+ years in academia, the two projects I am most proud of are STRIPS and Heterosis. The STRIPS project involved decades-long research plots to study the effect of planting strips of prairie within row-crop agriculture. The primary academic product was a PNAS manuscript that analyzed 44 response variables ranging from sediment runoff to weed cover to pollinator diversity. The pinnacle was inclusion of Prairie Strips in the Conservation Reserve Program as CP-43 in the 2018 US Farm Bill. The employment of my 11 descendants ranges from the University of Uruguay to Google. The Heterosis project, the PhD dissertation for one of these advisees, involved an RNA-seq experiment with 40k genes with 4 replicates of each of 4 varieties. Bayesian estimation of our hierarchical model, with over half a million parameters, required a custom MCMC algorithm written in CUDA C to run on a graphics processing unit (GPU). In our work, I hope that my students and I are making Mike proud.

Natesh Pallai
LinkedIn and Harvard University

Title. An unexpected encounter with the Cauchy distribution.
Abstract. We will discuss an elementary fact about the Cauchy distribution and present its proof. We will also present recent applications of this result in testing. Joint work with Xiao-Li Meng and Tyler Liu.

Abel Rodriguez
University of Washington

Title. Dirichlet mixtures of block g-priors for model selection and prediciton
Abstract. We introduce Dirichlet mixtures of block g-priors for model selection in linear models. These priors are extensions of traditional mixtures of g-priors that allow for differential shrinkage for various (data-selected) blocks of parameters. We show that Dirichlet block g-priors are consistent in various senses and develop a Markov chain Monte Carlo algorithm for posterior infer-

ence. In addition, we investigate the empirical performance of the prior in various real and simualted datasets. We conclude that, in the presence of a small number of very large effects, Dirichlet mixtures of block g-priors lead to higher power in detecting smaller but significant effects without substantially increasing the number false discoveries.

Yajuan Si
University of Michigan

Title. The Role of Data Collection in Population Science: A Bayesian Perspective

Abstract. The data landscape for population science research has recently shifted from using a single data collection mode to integrating multiple data sources. Without probability sampling or a randomization-based study design, the data quality and inferential validity require rigorous evaluations and may rely on untestable assumptions. Bayesian methods offer a natural solution to account for various sources of heterogeneity in data collection and to stabilize sparse group estimates through information sharing. I focus on the method of multilevel regression and poststratification and discuss the methodological challenges and opportunities in Bayesian approaches to improving data collection for population-based research.