SHORT COURSES

1) Bayesian Methods for Statistical Data Privacy - ROOM 10B
Instructors: Jingchen (Monika) Hu and Harrison Quick

Abstract:
Statistical data privacy is the field that deals with statistical methods for the public releases of confidential data, including record-level microdata, tabular data, summary statistics, and statistical test results. For the past few decades, statistical data privacy has generated much interest among researchers and practitioners. Since the beginning, Bayesian methods have played an important role in addressing data privacy challenges. These methods have ranged from synthetic data approaches that originated in the 1990s to the concept of differential privacy that arose in the 2000s. In this short course, we will cover 1) Bayesian methods for generating synthetic microdata; 2) for adding noise to tabular data to satisfy differential privacy leveraging conjugate models; and 3) creating differentially private synthetic datasets using Bayesian networks. The short course will provide ample opportunities for hands-on experience with synthetic data generation and noise infusion using R.

Suggested Prerequisites:
Participants will be assumed to have some familiarity with Bayesian inference and the R programming language. In particular, we will assume that participants are familiar with concepts such as Bayes Theorem, prior and posterior distributions, and simple conjugate prior models (e.g., beta-binomial and gamma-Poisson). In addition, we will assume that participants are comfortable with the basic functionality in R, including manipulating lists and arrays, using loops, installing packages, and generating random numbers.

Length:
Full day

Instructors' Biographies:
- **Jingchen (Monika) Hu** is an Associate Professor in the Department of Mathematics and Statistics at Vassar College. She teaches a senior seminar at Vassar on the topic of statistical data privacy and supervises undergraduate projects applying Bayesian techniques for creating synthetic data. Monika was an ASA Research Fellow at the U.S Bureau of Labor Statistics in 2018 working on synthetic data approaches to their survey products, a Faculty Fellow at the National Center for Science and Engineering Statistics from 2020 to 2021 working on differentially private synthetic data for tabular data products, and currently she is a consultant for New York City Department of Health and Mental Hygiene working on disclosure risk evaluation and mitigation solutions for several health surveys.
- **Harrison Quick** is an Associate Professor in the Division of Biostatistics at the University of Minnesota. In 2018, he was selected to serve as an ASA Research Fellow at the National Center for Health Statistics to conduct preliminary research on using Bayesian methods to generate differentially private synthetic data, work that led to an NSF CAREER award related to spatially referenced synthetic data. More recently, Dr. Quick received an R01 from NIH/NHLBI to develop Bayesian methods to help state and local health departments conduct small area analyses.

2) Sequential Monte Carlo: from state-space models to arbitrary posterior distributions - ROOM 9B
Instructor: **Nicolas Chopin**

Abstract:
This course will give a general overview of Sequential Monte Carlo methods (a.k.a. particle filters), and their different applications, from sequential inference in state-space (hidden Markov) models, to sampling from an arbitrary posterior distribution (and computing the corresponding marginal likelihood). I will try to cover many aspects, including a (gentle introduction to) Feynman-Kac distributions, methodological aspects, implementation (with examples in python), and applications in different areas.
This course is based on the following book: https://nchopin.github.io/books.html
and the python examples will rely on the accompanying library, particles:
https://github.com/nchopin/particles

Suggested Prerequisites:
The only prerequisites are probability (including a little bit of measure theory) and Bayesian inference. There is no need to be familiar with any form of Monte Carlo to follow this course.

Length:
Half day

Instructor's Biography:
- **Nicolas Chopin** (PhD, Université Pierre et Marie Curie, Paris, 2003) is a Professor of Statistics at the ENSAE, Paris, since 2006. He was previously a lecturer at Bristol University (UK). He is or was associate editor for Annals of Statistics, Biometrika, Journal of the Royal Statistical Society, Statistics and Computing, and Statistical Methods & Applications. He is a fellow of the IMS, and has served as a member (2013-14), and secretary (2015-16) of the research section committee of the Royal Statistical Society. He received a Savage award in 2002 for his doctoral dissertation. His research interests include computational statistics, Bayesian inference, and machine learning.

3) Introduction to optimal transport for Bayesian statistics - ROOM 9B
Instructor: **Hugo Lavenant**

Abstract:
Optimal transport is a flourishing theory enabling to build couplings between random objects and compute distances between probability distributions. With a rich historical background, this field has yielded significant contributions across probability theory, analysis, and geometry. Recent progresses on numerical methods have popularized it in the machine learning community. Its use is rising in Bayesian statistics with several important works regarding theoretical, methodological and computational aspects.

The aim of this minicourse is twofold: gaining a robust understanding of the theory and numerics of optimal transport while delving into its applications within Bayesian statistics. It will cover the fundamental notions and results of optimal transport, among which: Kantorovich formulation, Kantorovich duality, Wasserstein distances, algorithms for numerical optimal transport. In a second part it will illustrate a few more advanced topics with applications in Bayesian statistics, which could be, e.g.: Wasserstein barycenters for scalable Bayes, Wasserstein gradient flow to analyze continuous time sampling methods, optimal transport between completely random measures.

Suggested prerequisites:
No prior knowledge of optimal transport will be assumed. Familiarity with elementary concepts in probability and measure theory may be helpful.

Length:
Half day

Instructor's Biography:

- **Hugo Lavenant** is an Assistant Professor in the Department of Decision Sciences at Bocconi University (Italy) and a Research Affiliate to the Bocconi Institute for Data Science and Analytics. His research interests include optimal transport and the geometry of the Wasserstein space, convex analysis and optimization, calculus of variations and Bayesian statistics. He obtained his PhD in 2019 under the supervision of Filippo Santambrogio on the study of problems of calculus of variations involving Wasserstein distances. He later worked on problems with a more statistical flavor during his postdoc at UBC with Geoffrey Schiebinger and has recently done contributions on the use of optimal transport in Bayesian Nonparametrics.

4) Informative Prior Elicitation Using Historical Data - ROOM 2A
Instructors: Joseph G. Ibrahim and Ethan Alt

Abstract:

This full-day short course is designed to give biostatisticians and data scientists a comprehensive overview of informative prior elicitation from historical data, expert opinion, and other data sources, such as real-world data, prior predictions, estimates, and summary statistics. We focus both on Bayesian design and analysis and examples will be presented for several types of applications such as clinical trials, observational studies, environmental studies as well as other areas in biomedical research. The methods we present will be demonstrated in Stan and R.

The first part of the course will focus broadly on advanced methods for informative prior elicitation, including i) informative prior elicitation from historical data using the power prior (PP) and its variations including the normalized power prior, the partial borrowing power prior, the asymptotic power prior, and the scale transformed power prior (STRAPP). In addition, ii) the Bayesian hierarchical model (BHM) commensurate prior, and the robust meta-analytic predictive (MAP) prior will also be examined. The properties and performance of the four priors (BHM, PP, commensurate, robust MAP) will be analytically compared and studied via simulations and real data analyses of case studies. In addition, we will also examine iii) informative prior elicitation from predictions, including the hierarchical prediction prior (HPP), and the Information Matrix (IM) prior. We also examine iv) strategies for informative prior elicitation from expert opinion.

The second part of the course will focus exclusively on constructing informative priors for external control. Here, we discuss methods for both study design and analysis. We examine hybrid designs, propensity score methods for external controls, synthetic controls, meta-analytic methods, and power prior related methods for discounting external control data.

For (i) – (iv), we will present examples both in the context of Bayesian design and analysis and demonstrate the performance of these prior through several simulation studies and case studies involving real data in the context of linear and generalized linear models, longitudinal data, and survival data. We will also demonstrate the implementation of these priors through the R packages hdbayes, Nimble, and Stan.

Suggested prerequisites:

Students will be expected to have graduate knowledge of mathematical statistics and computing skills in the R programming language.

Length:

Full day

Instructors' Biographies:

- **Joseph G. Ibrahim**, Alumni Distinguished Professor of Biostatistics at the University of North Carolina at Chapel Hill, is principal investigator of two National Institutes of Health (NIH) grants for developing statistical methodology related to cancer, imaging, and genomics research. Dr. Ibrahim is the Director of the Biostatistics Core at UNC Lineberger Comprehensive Cancer Center. He is the biostatistical core leader of a Specialized Program of Research Excellence in breast cancer from NIH. Dr. Ibrahim's areas of research focus are Bayesian inference, missing data problems, cancer, and genomics. He received his PHD in statistics from the University of Minnesota in 1988. Dr. Ibrahim is a Fellow of the American Statistical Association

(ASA), the Institute of Mathematical Statistics (IMS), the International Society of Bayesian Analysis (ISBA), the Royal Statistical Society (RSS), and the International Statistical Institute (IMS).

- **Ethan Alt** is an Assistant Professor in the Department of Biostatistics and the Collaborative Studies Coordinating Center at the University of North Carolina Chapel Hill. He also serves as a statistical methodologist for the Center for Innovative Clinical Trials at UNC. Dr. Alt's areas of expertise broadly include the development of Bayesian methods for the design and analysis of clinical trials. He received his PhD in Biostatistics from the University of North Carolina at Chapel Hill in 2021.

5) Ensemble Learning with Bayesian Additive Regression Trees - ROOM 6A

Instructors: **Robert McCulloch**, **Rodney Sparapani** and **Antonio Linero**

Abstract:
Modern computing power has led to breakthroughs in our ability to learn high-dimensional, complex relationships from data. Recently, the two key modeling approaches in this arena are deep learning with neural networks and ensemble learning with trees. Deep learning is the best currently-known method of prediction where all of the covariates are of the same type, i.e., they are all pixels or words or audio waves, etc. Ensemble learning is the best currently-known method with respect to out-of-sample predictive performance for tabular data where all of the covariates are of different types, i.e., age, sex, weight, etc. A collection of machines (in our case trees) are fit simultaneously that form the basis of an ensemble's aggregate prediction with superior performance to any single machine's fit. In this workshop, you will learn a Bayesian approach to modeling with ensembles of trees called Bayesian Additive Regression Trees (BART). The Bayesian approach allows for a Markov chain Monte Carlo stochastic exploration of the model space, uncertainty quantification, and Bayesian posterior inference. BART is a modern nonparametric approach that exploits the elegance and convenience of the Bayesian conceptual toolkit. We can employ BART for outcomes of different types: continuous, dichotomous, categorical and time-to-event. Furthermore, we will demonstrate BART's effectiveness in a wide range of regression applications: marginal effects, variable selection, monotonicity, outlier detection and time-to-event extensions like competing risks or recurrent events. This is a lot of material to cover in one day. Therefore, we will be providing an overview with a wealth of materials for the attendees to self-explore afterward.

Suggested Prerequisites:
No prior knowledge of BART is assumed, but a passing understanding of Markov chain Monte Carlo posterior inference will be quite helpful. Comfort with basic R functionality is necessary including familiarity with regression interfaces (lm/glm/predict/etc.) and installing packages.

Length:
Full day

Instructors' Biographies:

- **Robert McCulloch** is a Professor in the School of Mathematical and Statistical Sciences at Arizona State University. He received his PhD from the University of Minnesota. Dr. McCulloch's research focuses on Bayesian statistics. The computational revolution has triggered a Renaissance of Bayesianism particularly in complex, high-dimensional data settings. Much of Rob's recent research is on Bayesian approaches to tree-based ensemble models known as Bayesian Additive Regression Trees (BART). BART has emerged as one of the most flexible tools in the Machine Learning toolbox. According to Google Scholar, he has over 7000 citations in the last 5 years including exemplar BART applications for personalized medicine, NHL penalties, selection of long-term portfolios, and scale conversion for marketing data. In 1997, he was elected as a fellow of the American Statistical Association.

- **Rodney Sparapani** has over thirty years of experience as a statistician mainly in academic research with stints in industry and government. He earned his PhD in Biostatistics from the Medical College of Wisconsin (MCW) working with his adviser Purushottam (Prakash) Laud. Rodney is now an Associate Professor of Biostatistics at MCW's Milwaukee campus. He was elected as chair of the ISBA Biopharm Section in 2021 and as the current President for the Wisconsin Chapter of the ASA. Last year, he was the winner of the ISBA Biopharm section's Best Paper Award for his article on NFT BART entitled "Nonparametric failure time: Time-to-event machine learning with heteroskedastic Bayesian additive regression trees and low information omnibus Dirichlet process mixtures" published in Biometrics (2023).

- **Antonio Linero** is an Associate Professor in the Department of Statistics and Data Sciences at the University of Texas at Austin. He obtained his PhD in Statistics from the University of Florida in 2015 advised by Mike Daniels and Hani Doss. Tony was a faculty member at Florida State University before joining UT Austin in 2019. His research focuses on developing Bayesian nonparametric methods for causal inference, missing data and high-dimensional problems. In this topic area, Dr. Linero has authored a book (with Daniels and Jason Roy) and has taught short courses at ISBA and ENAR meetings.