Invited sessions for ISCB42 (Lyon, 2021)
Session 1

Causal inference in continuous time for dense longitudinal data from wearable devices

Organizers : Linda Valeri : Columbia University Mailman School of Public Health, US

Mark van der Laan : University of California, Berkeley, US
TMLE for Causal Effects based on continuous time longitudinal data structures

Suchi Saria & Roy Adams , John C. Malone Associate Professor of Computer Science, Statistics and Health Policy and Director of the Machine Learning and Healthcare Lab, Johns Hopkins University , USA
The Impact of Time Series Length and Discretization on Longitudinal Causal Estimation Methods

Xiaoxuan Cai : Columbia University Mailman School of Public Health, USA
Missing data imputation for non-stationary time series in mHealth data

Title: TMLE for Causal Effects based on continuous time longitudinal data structures

Mark van der Laan[1] and Helene Rytgaard[2]

[1] UC Berkeley, USA - presenting author : laan@berkeley.edu

[2] University of Copenhagen, Denmark

Abstract:

In many applications one is concerned with estimation of the causal impact of a multiple time point intervention on a final outcome based on observing a sample of longitudinal data structures. We consider the case that subjects are monitored at a finite set of time-points on a continuous time-scale, and at these monitoring times treatment actions and or time-dependent covariates and outcomes are collected. Current methods based on sequential regression break down under this setting. We develop a new targeted maximum likelihood estimator that still avoids estimation of the conditional densities for outcome and covariates of likelihood, but instead estimates a conditional mean function. We also consider a TMLE that involves estimation of the conditional densities. We develop highly adaptive lasso estimators of the nuisance functions and establish asymptotic efficiency of the TMLE under minimal conditions. In particular, we demonstrate these new TMLEs for estimation of treatment specific survival functions for single time-point interventions on competing survival times. Advantages relative to first discretizing the time scale and using currently avaiable corresponding TMLE are discussed.

The Impact of Time Series Length and Discretization on Longitudinal Causal Estimation Methods

Roy Adams[1], Suchi Saria[1,2,3], Michael Rosenblum[4]

[1] Department of Computer Science, Johns Hopkins University, USA
[2] Department of Applied Mathematics and Statistics, Johns Hopkins University, USA
[3] Bayesian Health, USA
[4] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA
(presenting author)

The use of observational time series data to assess the impact of multi-time point interventions is becoming increasingly common as more health and activity data are collected and digitized via wearables, social media, and electronic health records. Such time series may involve hundreds or thousands of irregularly sampled observations. One common analysis approach is to simplify such time series by first discretizing them into sequences before applying a discrete-time estimation method that adjusts for time-dependent confounding. In certain settings, this discretization results in sequences with many time points; however, the empirical properties of longitudinal causal estimators have not been systematically compared on long sequences. In this talk, we compare three representative longitudinal causal estimation methods on simulated and real clinical data and analyze the impact of sequence length and discretization bin width on estimator performance. Our simulations and analyses assume a Markov structure and that longitudinal treatments/exposures are binary-valued and have at most a single jump point. We identify sources of bias that arise from temporally discretizing the data and provide practical guidance for discretizing data and choosing between methods when working with long sequences. Additionally, we compare these estimators on electronic health record data, evaluating the impact of early treatment for patients with a life-threatening complication of infection called sepsis.

Missing data imputation for non-stationary time series in mHealth data

Xiaoxuan CAI
Columbia University Mailman School of Public Health, USA

Missing data is a ubiquitous problem in studies of psychiatry, epidemiology, social, political science and many other biomedical and social science disciplines, when large number of variables are collected (especially with repeated measurements over time), and thus complete data are rarely available. As mobile devices (e.g., cell phones and fitness activity tracker bracelets) are being more widely adopted, a new way of collecting personal health data densely or even in real-time using mobile devices has realized, and revolutionized data collection methods for personalized health outcomes. Multivariate time series of outcomes, exposures, and covariates data evoke new challenges in handling missing data in order to get unbiased estimate of causal quantities of interest and call for more efficient data imputation approaches.

We conducted a comprehensive comparison of the performance of complete-case analysis with most commonly used imputation methods, including mean imputation, last-observation-carried-forward imputation, multiple imputation, multiple imputation with significantly longer history information, as well as state-space model in estimating causal quantities in mHealth data. Validity of most imputation methods rely on the stationarity of the time series, in the sense that both variance and treatment effect do not change over time, failing to reflect that the intervention effects in psychiatry may change with severity of disease as well as the fluctuation of the patient's mood. We propose a new imputation method derived from state space model, that accommodates the non-stationarity of time series when treatment effect and variance may change over time, to learn causal effects of interventions. We consider possible missing data in the outcome variable, exposure variable, or both under different missing rates and missing data mechanisms.