



Sumanta Basu
Department of Statistics and Data Science
Cornell University

Statistical Learning with Dependent Data

Wednesday, October 9th, 2024

12:00 PM

110 Frelinghuysen Road, Hill Center, Room 116

Zoom Meeting: Meeting ID: 969 0606 4706

Password: 745339

<https://rutgers.zoom.us/j/96906064706?pwd=ZklybExpRVBJQ3c5dUhhYTFuR2ZrZz09>

Light refreshments will be served in Hill 452, 11:20am

Abstract: With advances in data collection and storage, statistical learning algorithms are becoming increasingly popular for structure learning and prediction with large-scale data sets that exhibit temporal or spatial dependence. Most algorithms in the literature focus on using off-the-shelf machine learning algorithms that ignore the dependent nature of the data. In this talk, we aim to demonstrate the merit of incorporating classical statistical wisdoms for scale and dependence modeling into the statistical learning framework through two algorithms that we developed. The first, called RF-GLS, extends random forests (RF) for dependent error processes in the same way Generalized Least Squares (GLS) fundamentally extends Ordinary Least Squares (OLS) for linear models under dependence. The second algorithm, called AutoTune, offers an automatic tuning parameter selection algorithm for LASSO, by revisiting the well-known problem of scale estimation and adjustment for high-dimensional regression. We illustrate the benefit of these algorithms on simulated data sets, and provide some theoretical analysis to shed insight on their asymptotic properties.

Bio: I am broadly interested in developing statistical machine learning methods for structure learning and prediction of complex, high-dimensional systems arising in biological and social sciences. I am currently working in two areas: (a) network modeling of high-dimensional time series; and (b) detecting high-order interactions in complex biological systems using randomized tree ensembles. I also work closely with scientists and economists on a wide range of problems including prostate cancer progression, large scale metabolomics, and systemic risk monitoring in financial markets. My research has been supported by an NSF CAREER award (DMS-2239102), two three-year grants from the National Science Foundation (NSF DMS-1812128, DMS-2210675), a four-year grant (NIH R01GM135926) from Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences, and a two-year grant (NIH R21NS120227) from the National Institute of Health. Before joining Cornell, I was a postdoctoral scholar (2014-2016) in the Department of Statistics, UC Berkeley and the Biosciences Division, Lawrence Berkeley National Laboratory. I received my PhD (2014) from the Department of Statistics, University of Michigan, and my bachelors (2006) and masters (2008) in Statistics from Indian Statistical Institute, Kolkata. Before joining the PhD program, I worked for a year as a business analyst at Wipro Technologies.

