Data Science Seminar

Hosted by the Department of Mathematics and Statistics

- Date: Tuesday, October 15, 2024
- Time: 12:00pm – 1:00pm
- Room: Whitney Hall 100E
- Speaker: Dr. Yang Ning (Cornell University)
- Title: Estimation and Inference in Multivariate Response Regression with Hidden Variables.

Abstract

This paper studies the estimation of the coefficient matrix $\Theta $ in multivariate regression with hidden variables, $Y = (\Theta)^TX + (B^*)^TZ + E$, where Y is a m-dimensional response vector, X is a p-dimensional vector of observable features, Z represents a K-dimensional vector of unobserved hidden variables, possibly correlated with X, and E is an independent error. The number of hidden variables K is unknown and both m and p are allowed but not required to grow with the sample size n. Since only Y and X are observable, we provide necessary conditions for the identifiability of $\Theta $. The same set of conditions are shown to be sufficient when the error E is homoscedastic. Our identifiability proof is constructive and leads to a novel and computationally efficient estimation algorithm, called HIVE. The first step of the algorithm is to estimate the best linear prediction of Y given X in which the unknown coefficient matrix exhibits an additive decomposition of Θ and a dense matrix originated from the correlation between X and the hidden variable Z. Under the row sparsity assumption on $\Theta $, we propose to minimize a penalized least squares loss by regularizing $\Theta $ via a group-lasso penalty and regularizing the dense matrix via a multivariate ridge penalty. Non-asymptotic deviation bounds of the in-sample prediction error are established. Our second step is to estimate the row space of B∗ by leveraging the covariance structure of the residual vector from the first step. In the last step, we remove the effect of hidden variable by projecting Y onto the complement of the estimated row space of B∗. Non-asymptotic error bounds of our final estimator are established. The model identifiability, parameter estimation and statistical guarantees are further extended to the setting with heteroscedastic errors.

Biography of the speaker: Dr. Ning is an associate professor in the Department of Statistics and Data Science at Cornell University. He received his Ph.D in Biostatistics from the Johns Hopkins University and was a post-doc at Princeton University and University of Waterloo. His research interests focus on high-dimensional statistics, and statistical inference in machine learning problems.