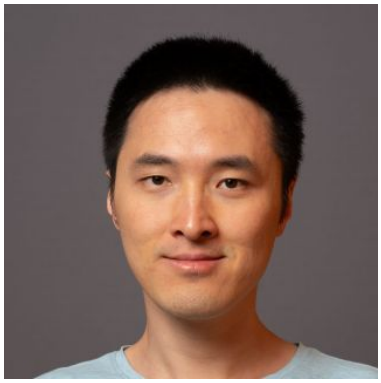
RUTGERS
School of Arts and Sciences

Department of Statistics

DEPARTMENT OF STATISTICS



Yang Ning
Department of Statistics and Data Science
Cornell University

Adaptive Estimation in Multivariate Response Regression with Hidden Variables

Wednesday, February 24, 2021
11:45 AM EST

Zoom Meeting: Meeting ID: 980 2426 8892
Password: 730799
https://rutgers.zoom.us/j/99537032528?pwd=ZXZGQnljVUJJaVVVVHhzd2dPRzBrUT09

Virtual Coffee session before the seminar at 11:30AM EST

Abstract: A prominent concern of scientific investigators is the presence of unobserved hidden variables in association analysis. Ignoring hidden variables often yields biased statistical results and misleading scientific conclusions. Motivated by this practical issue, this paper studies the multivariate response regression with hidden variables, $Y = (\Ps)^TX + (B^*)^TZ + E$, where $Y \in \RR^m$ is the response vector, $X\in \RR^p$ is the observable feature, $Z\in \RR^K$ represents the 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.
Though $\Ps$ is shown to be non-identifiable due to the presence of hidden variables, we propose to identify the projection of $\Ps$ onto the orthogonal complement of the row space of B^*, denoted by $\Ttheta$. The quantity $(\Ttheta)^TX$ measures the effect of X on Y that cannot be explained through the hidden variables, and thus $\Ttheta$ is treated as the parameter of interest. Motivated by the identifiability proof, we propose a novel %and computationally efficient estimation algorithm for $\Ttheta$, called HIVE, under homoscedastic errors. The first step of the algorithm estimates the best linear prediction of Y given X, in which the unknown coefficient matrix exhibits an additive decomposition of $\Ps$ and a dense matrix due to the correlation between X and %the hidden variable Z. Under the sparsity assumption on $\Ps$, we propose to minimize a penalized least squares loss by regularizing $\Ps$ and the dense matrix via group-lasso and multivariate ridge, respectively. Non-asymptotic deviation bounds of the in-sample prediction error are established. Our second step estimates the row space of B^* by leveraging the covariance structure of the residual vector from the first step. In the last step, we estimate $\Ttheta$ via projecting Y onto the orthogonal complement of the estimated row space of B^* to remove the effect of hidden variables. Non-asymptotic error bounds of our final estimator of $\Ttheta$, which are valid for any m,p,K and n, are established. We further show that, under mild assumptions, the rate of our estimator matches the best possible rate with known B^* and is adaptive to the unknown sparsity of $\Ttheta$ induced by the sparsity of $\Ps$. The model identifiability, estimation algorithm and statistical guarantees are further extended to the setting with heteroscedastic errors.

Bio: Dr. Ning is an assistant professor in the Department of Statistics and Data Science at Cornell University. Prior to joining into the Cornell University, he was a post-doc at Princeton University. He received his Ph.D in Biostatistics from the Johns Hopkins University. His research interests focus on the high-dimensional statistics and causal inference with applications to biology, medicine and public health.