THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

JUN LIU

Department of Statistics
Harvard University

Multiple Data Splitting for FDR Controls

MONDAY, November 2, 2020 at 4:00 PM
Via Zoom (session information will be e-mailed to subscribers)
Virtual Reception After Colloquium

ABSTRACT

Simultaneously finding multiple influential variables and controlling the false discovery rate (FDR) for regression models is a fundamental problem with a long history. Researchers recently have proposed the idea of creating "knockoff" variables (like spike-ins in biological experiments) to control FDR. As opposed to creating knockoffs, a classical statistical idea is to introduce perturbations and examine the impacts, such as bootstrap. We here examine how a simple and also old idea, data splitting (DS), can be leveraged for controlling FDRs. As one may have anticipated, a DS procedure simply estimates two independent coefficients, for each feature, from two datasets of the half-size created by random splitting, and constructs a contrast statistic. The FDR control can be achieved by taking advantage of the statistic's property that, for any null feature, its sampling distribution is symmetric about 0. Furthermore, via repeated sample splits, we propose *Multiple Data Splitting* (MDS) to stabilize the selection result and boost the power. Interestingly, MDS not only helps overcome the power loss caused by data splitting with the FDR still under control, but also results in a lower variance for the estimated FDR compared with all other methods in consideration. We prove that both DS and MDS can control FDR at the designated level asymptotically for canonical statistical models such as linear and Gaussian graphical models. DS and MDS are straightforward conceptually, easy to implement algorithmically, and also efficient computationally. They do not require any knowledge of the joint distribution of the covariates and are rather robust to certain violations of the theoretically required tail condition (e.g., t-distributions). Simulation results as well as a real data application show that both DS and MDS control the FDR well and MDS is often the most powerful method among all in consideration, especially when the signals are weak and correlations or partial correlations are high among the features. Our preliminary tests on nonlinear models such as generalized linear models and neural networks also show promises.

The presentation is based on joint work with Chenguang Dai, Buyu Lin, and Xin Xing.

For further information and inquiries about building access for persons with disabilities, please contact Jonathan Rodriguez at 773.702.8333 or send him an email at jgrodriguez@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/subscribe/statseminars.