THE UNIVERSITY OF CHICAGO

Department of Statistics

DISSERTATION PRESENTATION AND DEFENSE

ANG LI

Department of Statistics
The University of Chicago

Multiple Testing with Prior Structural Information

WEDNESDAY, May 24, 2017, at 1:00 PM
Jones 111, 5747 S. Ellis Avenue

ABSTRACT

Multiple testing problems arise when we simultaneously test thousands or even millions of hypotheses. In many applications, the hypotheses have certain structures, based on prior studies or domain knowledge, which is a valuable source of information. We study how incorporating such information could improve the performance of multiple testing.

Specifically, we first consider the ordered testing problem in which the hypotheses are ranked from the one most likely to be signal to the least likely one. Given this ordered list of n hypotheses, the goal is to select a data-dependent cutoff k and declare the first k hypotheses to be statistically significant while bounding the false discovery rate (FDR). Generalizing existing methods, we develop a family of "accumulation tests" to choose a cutoff k that adapts to the amount of signal at the top of the ranked list. Our theoretical results prove that these methods control a modified FDR on finite samples, and characterize the power of the methods in the family. We apply the tests to simulated data, including a high- dimensional model selection problem for linear regression. We also compare accumulation tests to existing methods for multiple testing on a real data problem of identifying differential gene expression over a dosage gradient.

We then introduce the structure-adaptive Benjamini-Hochberg algorithm (SABHA). SABHA incorporates prior information about any pre-determined type of structure within the list of hypotheses, to reweight the p-values in a data-adaptive way. This raises the power by making more discoveries in regions where signals appear to be more common. Our main theoretical result proves that SABHA controls FDR at a level that is slightly higher than the target level, as long as the adaptive weights are not overfit to the data—interestingly, the excess FDR is related to the Rademacher complexity of the class from which we choose our data-adaptive weights. We apply this general framework to various structured settings, including ordered, grouped, low total variation structures, and structures from community models, and get the bounds on FDR for each setting. We also examine the empirical performance of SABHA on fMRI activity, gene/drug response data, and on simulated datasets.

For information about building access for persons with disabilities, please contact Laura Rigazzi at 773.702-0541 or send an email to lrigazzi@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/arc/statseminars.