High Dimensional Hypothesis Screening Using P-value Perturbation

Ali Shojaie*
Department of Biostatistics, University of Washington, Seattle, WA USA ashojaie@uw.edu
Moulinath Banerjee
Department of Statistics, University of Michigan, Ann Arbor, MI USA moulib@umich.edu
George Michailidis
Department of Statistics, University of Michigan, Ann Arbor, MI USA gmichail@umich.edu

Abstract

High throughput biological experiments produce a large number of variables corresponding to activity levels of genes, proteins or metabolites in the cell. Statistical analysis of data from such experiments often start with an initial screening step to select "interesting" features for further analysis or followup experiments. A popular approach is to select variables through individual hypothesis tests, using the Neyman-Pearson inference framework, followed by multiple comparisons adjustment. In this paper, we propose an alternative method for screening high dimensional hypotheses using a perturbed version of regular p-values for two-sample inference. The proposed perturbation procedure is in line with Fisher's screening framework, and results in a dichotomous behavior in p-values for active and inactive hypotheses. Using the perturbed p-values, we subsequently develop a new procedure for selecting the set of active hypotheses. This framework alleviates the need for multiple comparisons adjustment and is shown to result in a consistent estimated of the set of active hypotheses in high dimensional settings, under arbitrary correlation structures.

keywords: hypothesis screening, multiple comparison adjustment, feature selection, correlated hypotheses