Data Integration of Multiple Genome-Wide Association Studies Under Group Homogeneous Structure

Yuan Jiang

Oregon State University
E-mail: yuan.jiang@stat.oregonstate.edu

Abstract: Nowadays, it's common to have a large collection of datasets or findings from similar scientific studies, with the famous example of multiple genome-wide association studies that are investigating the same human disease. To take advantage of these datasets or findings, statisticians have developed data integration methods to combine either raw data or summary statistics from multiple studies in order to increase statistical power. Most data integration methods to date can only combine compatible studies with the same explanatory variables; they also tend to ignore the grouping structure of the explanatory variables. However, incompatible studies with grouped explanatory variables arise frequently from multiple genome-wide association studies that employ different genotyping platforms. Therefore, we propose a new method called "gMeta" that can integrate incompatible raw data or summary statistics under a new group homoge- neous structure by utilizing group regularization principles. gMeta not only promotes statistical powers by assuming homogeneity among group-level signals but also allows heterogeneous individual-level signals from different studies. Simulation studies illus- trate the advantage of gMeta over separate analysis in terms of its consistency and enhanced statistical power for detecting weak signals. Finally, an integrative analysis of multiple genetic datasets on schizophrenia shows the applicability and efficacy of gMeta when it is applied to genome-wide association studies.