Joint Analysis of Imaging Genetics via Binary Matrix Factorization

Xing Xin, Xinlian Zhang, Cheng Meng, Wenxuan Zhong

The neuroimaging studies is revolutionizing our understanding of the neural pathways and mechanisms of the human brain through structure or functional imaging tectonics. In the meantime, large scale of genome-wide association studies unearthed numbers of genetic risk factors related to mental and neurological disorders. The imaging genetics, which combines the power of neuroimaging and genetics studies, is an emerging new field which aims at studying the association neuro-mechanisms and the linked genetic variation.

Comparing with the traditional genetic studies with illnesses or cognitive phenotype as a univariate response, the phenotypes in imaging genetics studies are usually a high-dimensional multivariate vector such as blood-oxygen-level dependent (BOLD) signals measured at different cortical or sub-cortical volumes. The imaging measure can localize brain activity to within millimeters, which provide an opportunity to study the association between the genetic factor and high-resolution region of the brain. The ultimate goal of the imaging genetic studies is to understand how the genetic factors affect the brain and further facilitate the diagnostic and cure treatment of brain disorders and illness.

The high dimension of both the genetic and imaging data rises new statistical challenges to study the association between the genetic factors and the brain. It is challenging to handle millions of predictors in the genome-wide association study. If we consider the two-way interactions between predictors, there will be more than 10^{11} predictors, which is much larger than the sample size. Current method can only deal with the case when $p < n$ or p is slightly larger than n. Besides, all existing methods assume all important predictors are observed, however many SNPs are missed in reality due to the limited sequence depth. When these assumptions are invalid, the estimation of parameters is biased.

We develop a binary component analysis (BCA) method to address these two aspects of the question. The predictor variables in genome-wide association studies, for example a nucleotide variations - C or A - at a base position having there combination CC, AA and AC, can be coded as two binary dummy variable. The dummy binary vectors are a subset of 2^{n-1} binary vectors. Comparing with the current variable selection methods, we search a larger space which contains all binary vectors instead of searching the space spanned by the observed predictor variables. Thus we not only include both the observed and missed predictor variables but also take into consider all the interactions. The major goal is to select minimum number of vertexes that can cover the space spanned by the response Y. Thus we propose a fast algorithm projecting Y on a low dimensional space then searching for a minimum number of binary vectors that can cover this projected space. The algorithm is motivated by the binary matrix factorization algorithm which is arise in blind source separation, inmixing of cell mixtures from DNA methylation signatures. We further refine the estimation by using alternating least square algorithm. The proposed method is compared with the existing rank reduced regression and regularized regression models in simulation study and shows competitive performance.