Functional Genome Wide Association Studies Using Sparse Group Lasso

Tatiyana Apanasovich
George Washington University, Washington DC, United States apanasovich@gwu.edu

Qing Pan*
George Washington University, Washington DC, United States qpan@gwu.edu

With the rapid development of efficient and inexpensive high throughput genotyping techniques, Genome Wide Association Studies (GWAS) provides powerful tools that map over a million single nucleotide polymorphisms (SNP) covering the complete genome. The commonly used one SNP at a time testing strategy fails to take into account the effect of multiple SNPs simultaneously and is subject to severe multiple comparison adjustments. Furthermore in complex diseases, the genetic effects are usually weak and changes over time. In cases of phenotypic data collected at various time points, single time point methods do not utilize information available in data effectively and may miss important genetic associations. Hence the models which incorporate the dynamic pattern of genetic influences over a time course are in a great need.

In this article we develop a novel nonparametric model that incorporates the association of multiple SNPs simultaneously with a continuous longitudinal trait of interest measured at irregularly spaced time points, which are not common to all subjects. Therefore, the effects of covariates and SNPs are assumed to be functions of time. Specifically, we consider B-spline polynomials to approximate time varying effects and propose to use sparse group lasso penalty, which introduces both group-wise and within group sparsity. The selection of important SNPs corresponds to the selection of groups of spline coefficients.

The statistical properties of the proposed method are investigated through simulation studies. We apply out method to the GWAS study of the Epidemiology and Intervention of Diabetes Complication trial where Type 1 Diabetes patients are followed for up to 27 years and annual glomerular filtration rates (GFR), a clinical measure for nephropathy, are taken.

Key words: Diabetes complications; Functional data analysis; B-spline smoothing; Variable selection