Partial-Least-Squares-Regression models for genome-wide association studies

Garrett Hellenthal[1] and Christopher C. Holmes[1]

[1] Department of Statistics, University of Oxford, UK

The power of genome-wide association (GWA) studies to detect associations between genetic variants and disease susceptibility has been amply demonstrated in the recent series of papers detailing the discovery of significantly-associated variants that have successfully replicated (e.g. [1, 2]). Recent technological advancements in genotyping have led to drastic decreases in the cost of collecting individual genetic variation data, in particular Single-Nucleotide-Polymorphism (SNP) data. As a result, typical GWA studies currently include the genetic information from hundreds of thousands of SNPs collected in thousands of individuals, with these numbers likely to increase even further over the next few years. Such studies bring new challenges in statistical analysis, including the need to develop methods that can deal efficiently with the vast amounts of data and account for the complex correlation structure amongst SNPs across the genome. Dealing with the potentially high correlations amongst SNPs in densely genotyped regions is likely to become even more of a prominent factor as studies shift from collecting less correlated "tag-SNP" data (where SNPs are often selected to have, e.g., pairwise correlation coeffients below some threshold) to exhaustive "resequencing" data (where all SNPs amongst the cohort individuals are collected).

One recently proposed means to address the correlation amongst SNPs advocates the use of Principal-Components-Regression (PCR) analysis [3]. Analagous to simple linear regression, PCR tests for associations between phenotypes and linear combinations of SNPs. These linear combinations are selected in a manner to explain as large a proportion of the genetic variability in a given region as possible while using a parsimonious representation of the data, thus both accounting for correlated SNPs and down-weighting "uninformative" SNPs.

We have developed an alternative technique that utilizes Partial-Least-Squares-Regression (PLSR) analysis. PLSR selects linear combinations of SNPs in a manner that attempts simultaneously to explain as much of the genetic variability in a given region and as much of the covariance between the SNPs and phenotypes as possible. It seems probable that PLSR might perform better than PCR in a GWA setting, as it may avoid down-weighting SNPs that are associated with the phenotype of interest but account for little of the genetic variance. Numerous algorithms exist to efficiently implement PLSR so that it is easily applicable to the scale of current studies. We explore a variety of simulations to test the potential benefits of PLSR, in particular a new PLSR technique that includes features from Bayesian Regression. Our results suggest that the technique can offer vast improvements in precision and power to detect disease-susceptibility variants over the most widely-used current methods in "resequencing" settings, while suffering no loss of power in "tag-SNP" settings.

References

[1] The Wellcome Trust Case Control Consortuim (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678

[2] Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331-1336

[3] Wang K, Abbot D (2008) A principal components regression approach to multilocus genetic association studies. Genetic Epidemoiology 32:108-118