Penalized nonlinear canonical correlation analysis for whole genome association studies

Sandra Waaijenborg[1] and Aeilko H Zwinderman[1]

[1] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, NL

Gene expressions are regulated in complex pathways that are influenced by many extra- and intracellular factors among which are expression of other genes, gene copy numbers, DNA methylation, and DNA structure using single nucleotide polymorphisms (SNPs). All of these factors can be and are in practice determined in many biological and epidemiological areas to dissect the heritable component of complex traits. In whole genome association studies these factors are measured in large numbers, and the challenge is to quantify association between expression and these other factors. We focus on the association between whole genome gene expression and whole genome SNPs.
In this paper we describe a penalized nonlinear canonical correlation approach for quantifying the genotype-expression associations that takes account of the multinomial character of the SNP-data by estimating the quantification of the wild type, heterozygous, and homozygous genotypes. We use the elastic net [1] as a penalty function to limit the number of gene expressions and the number of SNPs that are used to calculate the canonical components, and this improves interpretation of the results. Penalty parameters are estimated with 10-fold crossvalidation. Since the number of gene expressions and the number of SNPs is usually very large, the canonical correlations are necessary close to unity in sample data, even if in the population it is zero. We use permutations to distinguish between such a situation and a true high canonical correlation.
Using simulations we illustrate that our approach is capable to identify clusters of genes and SNPs that are truly associated, and we finally present an application of our method on whole genome gene expression (54.613 genes) and SNP (53.986 SNPs) data of 144 patients with glial cancer [2].

References

[1] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J. Royal Statistical Soc. Ser. B; 67: 301-320

[2] Kotliarov Y, Steed ME, Christopher N, Walling J, Su Q, Center A, Heiss J, Rosenblum M, Mikkelsen T, Zenklusen JC, Fine HA (2006) High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. Cancer Res; 66(19):9428-36