ABSTRACT

In genome-wide association studies (GWAS), it is of interest to infer population structure in the sample as false-positive results can arise due to population heterogeneity in terms of ancestry. A common approach is to use principal component analysis (PCA) that summarizes genetic similarity between subjects at tens of thousands of single-nucleotide polymorphisms (SNPs). However, it may be sufficient to use only a small number of SNPs, that we refer to as ancestry-informative markers (AIMs), in order to infer population structure. A novel technique, based on sparse PCA, is proposed which uses penalized regression methods to identify the set of AIMs from the genome-wide SNP data. This talk is based on the paper by Lee et al. (2012).