THE UNIVERSITY OF CHICAGO

Department of Statistics

DISSERTATION PRESENTATION AND DEFENSE

SHENG ZHONG

Department of Statistics

The University of Chicago

Mixed-Model Methods for Genome-Wide Association Analysis with Binary Traits

WEDNESDAY, July 2, 2014, at 2:00 PM

117 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Genome-wide association analysis has been widely applied for the last decade, as a major tool to identify the risk variants contributing to complex diseases. For association mapping of quantitative traits, linear mixed models (LMMs) have been used to account for relatedness and population structure as well as covariates. Such methods can be applied to binary traits as well. However, the use of a linear mean structure can result in a loss of power in this case. Use of a generalized linear mixed model (GLMM) can avoid that problem, but the resulting methods can be very computationally challenging in the context of genome-wide association studies in samples with related individuals.

In this thesis, we develop a new case-control association testing method, MABRAC, which is applicable to samples with related individuals and is scalable for genome-wide analysis. MABRAC is based on a novel estimating equation approach that can be viewed as a hybrid of logistic regression and LMM approaches. In terms of statistical power, MABRAC outperforms or performs as well as approaches based on either logistic regression or LMMs. Our method results in an even greater power increase over the other methods when the sample includes missing data (genotype, phenotype or covariates) and also has related individuals. In that case, we make full use of the relationship information to incorporate partially missing data in the analysis while correcting for dependence. Unlike available GLMM-based methods, our estimating equation approach is computationally efficient and stable for genome-wide analysis. Furthermore, MABRAC performs retrospective analysis, and so is robust to the misspecification of the phenotype model, because the correct calibration (i.e. type 1 error) of MABRAC only relies on the null variance of the genotype data. MABRAC is applicable to samples with arbitrarily-related individuals, including samples that combine small families and unrelated individuals as well as individuals related by large, complex inbred pedigrees, assuming that the genealogy is known.

We conduct simulation studies to confirm the correct calibration of MABRAC and its superiority in terms of power over other methods, and we apply MABRAC to analyze data on type 2 diabetes from the Framingham Heart Study.
