Shape mapping in genome-wide association studies

Xiaotian Dai
Dept. of Math and Statistics
Utah State University
1111 N 800 E Apt B404
Logan, UT 84341
Email: xiaotian.dai@aggiemail.usu.edu

Keywords:
GWAS
Shape mapping
Functional data

Background
Genome-wide association studies (GWAS) interrogate large-scale whole genome to characterize the complex genetic architecture for biological traits. Currently, the major focuses of GWASs are the associations between single-nucleotide polymorphisms (SNPs) and traits such as human diseases. We are particularly interested in the associations between SNPs and biological shapes like leaf shapes. This is an interdisciplinary research involving geometrical analysis and statistical shape and genetics analysis, which also face some unprecedented challenges.

Problem Statement
In the scenario of big data, there are two major obstacles in shape genetic mapping studies.

Firstly, rapid progress in genotyping technology is generating large numbers of SNPs and inspiring new findings in biological studies. When the number of SNPs dramatically increases to millions but the sample size is still limited to thousands, the traditional p-value based statistical approaches suffer from loss of power. Feature screening has proved to be an effective and powerful approach to handle ultrahigh dimensional data statistically, yet it has not received widely enough attention in GWAS. In addition to the large number of SNPs, the univariate measures used to rank features are mainly based on individual marginal effect without considering the mutual interactions with other features. It is very computationally expensive to include every possible combination of interaction effects into a regression model as the number of predictors (SNPs) becomes very large, which can also easily render traditional regression methods to lose power.

Secondly, it is impossible to represent biological shapes using a univariate response variable as in traditional regression models. Some geometric models can transform leaf
shapes or other biological shapes into functional data based on their geometrical measurements, such as the radius-centroid-contour (RCC) model discussed in Fu et al. (2013). According to the RCC model, we can assign points on the boundary of a leaf or other organisms with equal radial angle $\theta$, where $\theta = 2\pi/n$ and $n$ is the number of points assigned. Then we can quantify the shape as a functional RCC curve defined as a radius function of radial angle $\theta$ at the centroid. The whole process of shape extraction is illustrated in Figure 1. However, many of the traditional genetic mapping methods cannot handle high-dimensional responses. One of the mainly used approaches in other researches is principal component analysis (Ziezold, 1994). Nevertheless, it would be more innovative if the functional data representing shapes can be directly included into a regression model as responses.

![Figure 1. The procedure of extracting shape information from a leaf image. The RCC curve (D), as a function of the radial angle, can uniquely represent a shape (A).](image)

These problems described above not only are the obstacles in genetic studies, but also are challenges in practical data science problems. It would be very promising to develop a new machine learning algorithm to handle high-dimensional response and predictor space simultaneously.

**Broader Impacts**

As mentioned above, the major motivation of this white paper is to develop a machine learning algorithm to handle high-dimensional response and predictor space simultaneously. This algorithm can be applied to genetic mapping studies. Nevertheless, it has many other applications in a wide range of practical problems. In the scenario of big
data, the existing statistical models face many unprecedented challenges due to the complexity of both the data collected and the problems we want to solve.

**References**
