Multivariate statistics and population structure in large genetic datasets

Main supervisor

Dr James Bentham

School of Mathematics, Statistics and Actuarial Science (SMSAS), University of Kent, UK

j.bentham@kent.ac.uk

Aim: The student will use a range of multivariate statistical methods to quantify population structure in genetic data. At present, principal component analysis (PCA) is the standard method. Principal components are then used as covariates in subsequent regressions. The student will investigate alternatives to PCA, with the aim of increasing power to detect disease-causing genetic variants.

Background: Huge volumes of genetic data have been generated over the past ten years. The aim of much of this work is to find genetic variants that increase the risk of developing particular diseases.

One common type of genetic experiment is the genome-wide association study. Genetic samples are taken from large numbers of people with a disease, and even greater numbers of healthy controls. The samples are used to generate data for several hundred thousand or more *single-nucleotide polymorphisms* (or SNPs) for each person. The data are counts of the number of variant forms each person has at each SNP (either 0, 1 or 2). Regressions are then fitted to identify SNPs where frequencies vary significantly between individuals with the disease and healthy subjects.

These analyses can be confounded by population structure, i.e., systematic differences between groups in SNP frequencies that are not related to disease. Normally, principal components are used as covariates to correct for this structure. However, alternative multivariate statistical methods might provide improvements in statistical power in subsequent regression analyses.

Student's expected contribution: The student will use data downloaded from public repositories and will also simulate data using existing programmes. The student will need to pre-process the data carefully. They will then investigate the effect of applying various multivariate statistical methods to the genetic data. This will allow them to assess how regression analyses behave using the different methods. Should an alternative method prove to be beneficial, the student is likely to produce an R package or similar output. This would allow efficient application of the method by other researchers. A student who is particularly interested in multivariate statistics might choose to extend an existing method, or even develop something new. The size of the datasets will pose a key challenge that the student will need to overcome.

Skills required: The student will require an MSc in Statistics or equivalent, and a strong interest in multivariate statistics. The project will involve the analysis of very large datasets, with the work being carried out using Linux. There are various statistical genetics programmes run from Linux that other researchers have developed, and which the student might use. The student is also likely to program using R but may need to use an efficient language such as C++. While the student could learn how to use Linux and/or C++ during the project, prior knowledge would be beneficial. The student should also be interested in the analysis of genetic data.

The student will be based in SMSAS at the University of Kent and will be assigned a second supervisor from the school.