Ridge regression for risk prediction with applications to genetic data

Technological developments in high-throughput genotyping have increased the availability of genetic data. Current studies genotype hundreds or thousands of individuals at hundreds of thousands to millions of genetic variants (typically single nucleotide polymorphisms, or SNPs), Genotyping costs continue to decrease, and it is likely that future studies will incorporate sequence data and thus even greater numbers of genetic variants.

One potential application of these data is the identification of genetic variants associated with disease phenotypes. Identified variants could indicate potential novel drug targets or give insights into a disease mechanism, leading to suggestions of new pharmaceutical approaches to investigate.

Data obtained in large-scale genetic studies can also be applied to the development of models for the prediction of disease risk. Because genotypes are fixed at birth, but disease symptoms may not present until later in life, risk prediction using genetic data could suggest lifestyle or pharmaceutical interventions in higher risk individuals, reducing future disease susceptibility.

The use of genetic data to construct predictive models presents some statistical challenges due to the high dimensionality of the data and the correlation among nearby SNPs. Traditional multiple regression fails under these circumstances. Current studies that aim to incorporate genetic factors into risk prediction models use a small number of genetic variants most strongly associated with disease risk. However, the resultant risk prediction models offer little or no improved predictive ability compared to risk prediction models using clinical risk factors alone. In complex diseases, it is plausible that a much larger number of genetic variants contribute, and this prior belief is accounted for in the method used in this study.

We apply ridge regression to the development of models for risk prediction using genetic data. Ridge regression is a penalized regression method that gives rise to stable parameter estimates even when the number of predictors exceeds the number of observations, as is typically the case for genetic data. From a Bayesian perspective, coefficients estimated using ridge regression are the maximum *a posteriori* estimates when the prior distribution is Gaussian. The prior variance is controlled by the ridge parameter. A number of methods have been proposed in the literature to choose this ridge parameter based on the data, but existing studies focus on minimising mean squared error of the resultant parameter estimates, and no consensus method provides a universally optimum choice. In this study we investigate a means of choosing the penalty parameter such that the degrees of freedom for variance is the same as that of a principal components regression with a specified number of principal components. We discuss ways to choose the number of components to use, with the aim of good predictive performance. We demonstrate that when the number of causal variables is large and effect sizes are small, a plausible situation in the case of complex diseases, our method offers improved predictive performance over other regression methods. We apply our method to out-of-sample prediction using two Bipolar Disorder genome-wide association studies.