Building gene expression profiles via multivariate analyses

Manuela Zucknick[1] and Sylvia Richardson[1]
[1]Centre for Biostatistics, Imperial College London.
St. Mary's Campus Norfolk Place, London W2 1PG, UK

One application of gene expression arrays is to derive molecular profiles, i.e. small sets of genes, which discriminate well between two classes of samples, for example between two types of tumours or between favourable and unfavourable disease outcomes. Finding the molecular profiles is a variable selection problem in the "large p, small n" situation where the number of variables is much larger than the sample size. Such problems are ill-conditioned and high multi-collinearity among covariates implies that there might not be one unique best variable subset, but many equally good solutions. Thus, the question arises how stable molecular profiles are, that is how much they vary when there are slight changes in the training data.

A standard approach to building molecular profiles is to select genes based on some univariate measure such as correlation with the response (univariate filtering methods). However, this is not a recommended variable selection strategy, because the correlation structure among covariates is not accounted for. Here, we investigate molecular profiles derived from multivariate sparse penalised likelihood methods, in particular ridge and LASSO regression. We assess the stability of the profiles by performing a multiple random validation study and compare them with profiles derived from the standard univariate technique. We apply the methods in a logistic regression framework and perform a case study on well-established cancer gene expression data sets.