Colouring and breaking sticks, pairwise coincidence losses, and clustering expression profiles

Peter Green
Bristol University

We consider methodology for Bayesian model-based clustering of gene expression profiles, that is, measurements of expression levels of a large number of genes, typically from microarray assays, across a number of different experimental conditions and/or biological subjects. We follow a familiar approach using Dirichlet-process-based models to cluster the genes implicitly, but depart from standard practice in several ways. First, we incorporate regression on covariate information at the condition/subject level by modelling regression coefficients, not the expectations of the data directly. More importantly, we replace the Dirichlet process by one of a richer family of models, generated from a stick-colouring-and-breaking construction, under which cluster identities are not exchangeable: this allows modelling a 'background' cluster, for example. Finally, we follow a formal decision-theoretic approach to point estimation of the clustering, using a pairwise coincidence loss function. This is joint work with John Lau, previously at Bristol.