A Bayesian hierarchical sparse factor model for complex experiments in genetical genomics

Genome-wide gene expression data can provide rapid insight into mechanisms underlying responses to genetic variants or environmental perturbations. Gene expression datasets generated from complex multi-factor experiments are increasingly common in medicine, evolutionary biology, and agriculture. Robust models for identifying sets of genes associated with particular factors in experiments with correlated samples are lacking. Here, we propose a Bayesian model that aims to uncover gene expression signatures of the response to particular experimental treatments, such as plant density, in experiments with design features like sub samples, split plots or repeated measures, or with the experimental units sharing an arbitrary covariance such as from a pedigree or a structured population. We assume that genes function within semi-independent modules (factors), and that these factors are latent traits that vary according to the experimental treatments, genetic backgrounds, and any additional micro-environmental variation. This implies a factor structure for the gene expression covariation that is highly structured – both across genes and among samples – and that can be characterized by a relatively low number of parameters, which we enforce with biologically-motivated sparsity-inducing priors. To account for sub-samples or other pseudo-replication in the experimental design, we employ an efficient hierarchical mixed effect model simultaneously for the observed gene expression traits and the underlying latent traits. The advantages of this approach are two-fold. First, the mixed effect model permits modeling of the raw data, rather than on means of subsamples, and so can leverage the among gene correlation structure. Second, the factors themselves can be explored to provide biological intuition into the mechanisms driving biological responses to the experimental factors by inspecting functional or pathway classifications of genes in the modules. We demonstrate our approach on a large RNAseq dataset from *Arabidopsis thaliana*.