

Multi-level mixture modeling and Simultaneous subset selection, with applications to clustering and significance analysis of gene expression data

Dr. Rebecka Jornsten,
Assistant Professor
Rutgers University – Busch Campus

ABSTRACT

The analysis of gene expression data present many challenges that can be formulated as model selection problems. In model-based clustering, we group genes that exhibit similar expression profiles across experimental conditions. To allow for direct and objective inference of the clustering outcome, we need to determine a sparse representation of each cluster; between which experimental conditions does the cluster expression profile truly differ?

Model selection in clustering is combinatorial in the number of clusters and the number of experimental conditions, and thus presents a computationally challenging task. We introduce a simultaneous approach to subset model selection, which draws on results from rate-distortion theory.

The rate-distortion formulation allows us to turn the combinatorial model selection into a fast and simple line search. Furthermore, by considering each gene as its own cluster, the simultaneous selection framework extends to significance analysis of differential expression. We can thus determine not only if a gene is differentially expressed, but also which are the discriminatory experimental conditions.

These days, data often have a complex structure, and the clustering techniques we apply should reflect this. We introduce multi-level mixture models to address this issue. The multi-level framework can incorporate multiple distance metrics into clustering simultaneously, and be used to analyze multi-factor experiments. Multi-level mixture models extend model selection in clustering to between-cluster comparisons, and can constitute a substantial savings of model parameters, allowing for more clusters to be detected than with standard clustering techniques.