High-Dimensional Analyses of Gene Expression Data

Ernst Wit (ernst@stats.gla.ac.uk)
Department of Statistics, University of Glasgow

Abstract

One of the challenges that statisticians face is to perform biologically relevant inference within the scope of what is computationally feasible. We present two alternative ways to deal with biological questions in a high-dimensional stochastic system. The first is a fully probabilistic modelling approach, in which it is of interest to detect spatial patterns of gene expression interactions. The second is a data mining method based on "partitioning by medoids".

1 Hidden Markov Modelling of Gene Interactions

Several biologist have suggested that gene expressions may have a spatial component on the genome. The type of spatial component we consider are probabilistic nearest-neighbour interactions between the hidden states: being under-, not and over-expressed. The data set consists of a time-series (4 time points) of the M. Tuberculosis, which has, like most bacteria, a circular genome. We define a Hidden Markov Random Field (4×3924), where the genes represents the columns and the rows represent the time points. For each element we observe a gene expression. Figure (a) below give a graphical representation of the hierarchical model. Interactions are defined on the hidden states s, which in its simplest form can be written as

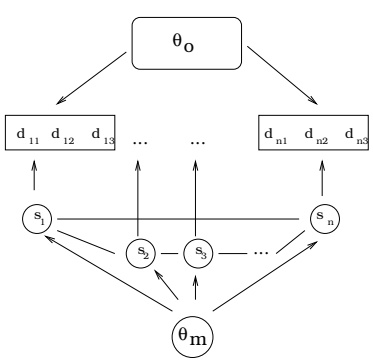
$$p(s_{ij}|s_{-ij}\ \theta_m) \propto \exp\left(\theta_t \sum_{m=i-1}^{i+1} \frac{2-|s_{mj}-s_{ij}|}{2} + \theta_g \sum_{n=j-1}^{j+1} \frac{2-|s_{in}-s_{ij}|}{2}\right),$$

where θ_t is the time interaction coefficient and θ_g is the genome interaction coefficient. We use a hybrid Gibbs and Metropolis-Hastings sampler to estimate the parameters. Typically this would involve pseudo-likelihood due to the presence of an essential normalizing constant. However, we use an exact method due to Pettitt, Reeves and Friel (2002) that avoids this approximation.

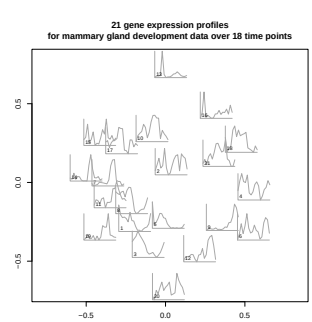
2 PAMSAM: Multi-dimensional scaling of cluster medoids

An extremely useful parameter in clustering is that of the average silhouette width (ASW), due to Kaufman and Rousseeuw (1990). Maximizing ASW in algorithms, leads to very sensible clustering results. The same people also devised "partitioning around medoids" (PAM), a form of k-medians. However, this algorithm does not use ASW to define clusters. Also, since PAM works on the $n \times n$ distance matrix, PAM has a physical limit to the number of points it can cluster. On the other hand, PAM has good clustering qualities, though splitting large clusters more than needed.

We define an algorithm that combines the flexibility of a sampling approach (CLARA) and the good clustering properties of PAM to define an iterative clustering algorithm that aims to maximize the ASW for large data sets. The result of the clustering is presented as a Sammon plot of the medoids, such as Figure (b) for a mammary gland development data set.



(a)



(b)

Acknowledgments: This is joint work with Nial Friel and John McClure (Statistics, Glasgow).