



Abstract

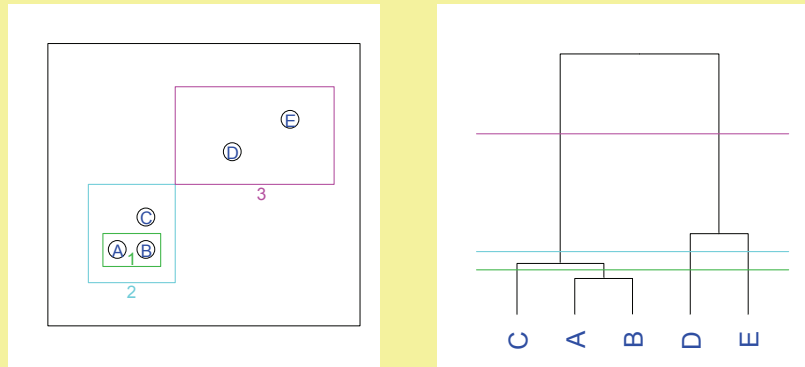
Statistical clustering and variable selection are important in many applications. In high-dimensional settings such as metabolomics or cDNA data, the structure is usually hidden in small subsets of variables, selection of which while clustering is burdensome. The increasing number of applications requiring this has stimulated the search for fast, automatic and accurate clustering algorithms for noisy data. Here we present a Bayesian agglomerative clustering algorithm based on a spike and slab model, used to kill unnecessary high dimensions when clustering. The use of the method is presented on replicated metabolomic, and on cDNA cancer data. The results are plausible and algorithm is much more rapid than other competing approaches.

Introduction

Clustering is the science of finding groups in data. Classical clustering techniques are mainly based on a measure of similarity or a distance, and often the result is shown in a tree-shaped graphic called a *dendrogram*. From a statistical viewpoint, clustering may also be regarded as fitting a mixture model with an unknown number of components. Bayesian model-based clustering is fashionable nowadays, partly due to the availability of Markov chain Monte Carlo algorithms that aid the computation of the posterior distribution. Finding all possible groupings of data is computationally too expensive with a lot of subjects, since the number of possible groupings grows rapidly as the number of subjects increases, so agglomerative clustering is often used to get results rapidly.

Dendrogram Construction in Agglomerative Clustering

The word agglomerative refers to the method of construction of a cluster. Based on a similarity measure it starts with every subject as a different cluster and successively merges the closest clusters, ending with all subjects in one cluster. The height of the dendrogram is given by the similarity measure at that step of clustering. Cutting the tree at different heights leads to different groupings.



Hierarchical Bayesian Model

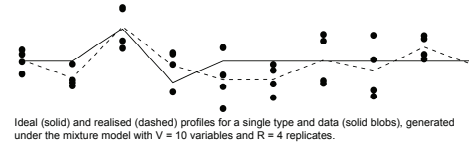
We consider data in which substances of different types $t = 1, \dots, T$ are processed in slightly different experimental settings $s = 1, \dots, S$. There are R_{st} replicates of type t under setting s , for non-replicated case $R_{st}=1$. The data for each replicate consist of observations y_{stvr} where the index for the variables varies in the range $v = 1, \dots, V$. The result of experiment is y_{stvr} with overall mean μ may be expressed as

$$y_{stvr} = \mu + \eta_{st} + \gamma_{tv} \theta_v + \varepsilon_{stvr}, \quad v=1, \dots, V, \quad t=1, \dots, T, \quad s=1, \dots, S, \quad r=1, \dots, R,$$

where θ_v , η_{st} and the between replicate error ε_{stvr} are independent continuous random variables with means zero and respective variances $\sigma_{\theta_v}^2$, σ_{η}^2 and σ^2 , and γ_{tv} is a binary random variable with success probability p . We suppose that type-variable combinations are active independently with probability p and otherwise are inactive. The indicator variables γ_{tv} indicate whether or not each variable-type combination is active. In the case of activation, $\gamma_{tv}=1$, introduces a standard density (the slab) for the effect, otherwise it introduces a point mass at 0. The profile mean under ideal experimental conditions is θ_v , which is taken to have a Gaussian distribution. If a combination is inactive, then the profile mean for that variable-type pair takes a baseline value μ .

The model induces η_{st} , as experimental setting variation, the difference between the ideal profile and that actually realised. Observed data y_{stvr} is normally distributed around the realised profile. We may rewrite the model in hierarchical form as

$$\begin{aligned} y_{stvr} &| \eta_{st}, \sigma^2 && \stackrel{i.i.d.}{\sim} N(\eta_{st}, \sigma^2), \\ \eta_{st} &| \theta_v, \sigma_{\eta}^2 && \stackrel{i.i.d.}{\sim} N(\theta_v, \sigma_{\eta}^2), \\ \theta_v &| \gamma_{tv}, \sigma_{\theta}^2 && \stackrel{i.i.d.}{\sim} N(\mu, \gamma_{tv} \sigma_{\theta}^2), \\ \gamma_{tv} &| p && \sim \text{Binom}(1, p), \end{aligned}$$



where $N(\mu, \sigma^2)$ represents the Gaussian distribution with mean μ and variance σ^2 , and $\text{Binom}(1, p)$ denoted the Bernoulli distribution with success probability p . The hyperparameters may be estimated by directly maximizing the marginal likelihood of the model, which has a closed form. We drop the index s later on, due to assumption of having one experimental setting for each profile. When the proportion of true effects equals to zero, every profile would be independent of every other, thus reducing over-fitting by the cluster algorithm when p is small.

Bayesian Agglomerative Clustering

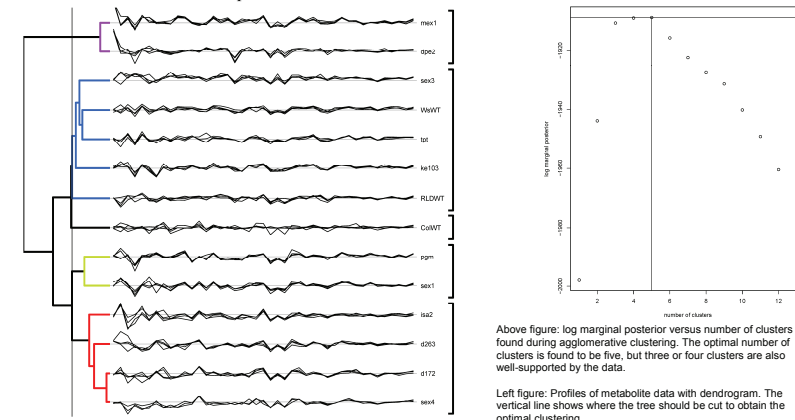
Unlike traditional methods which use distance to find adjoining clusters, we use the log marginal posterior as a measure of similarity. To define partitioning we may use a vector of integers called *labels*, with T elements that allocate types to clusters. In order to calculate the Bayesian posterior of the clustering, we use a uniform discrete distribution for the number of distinct clusters and a uniform multinomial-Dirichlet distribution for the labeling given the number of clusters. As a consequence of the spike and slab model when $p=0$, the posterior probability of any partition equals the prior probability, and the most probable clustering *a priori* is a single cluster containing all types.

At each step of the agglomerative clustering algorithm, every possible merge of pairs of blocks is considered, and the merge maximising the posterior probability is applied. The height of dendrogram in that step is taken to be the log likelihood of the posterior at the merge.

For graphical representation of the agglomerative clustering a monotone height is needed. We subtract the maximum value and replace heights by negative values when passing through zero. Then cutting the dendrogram at height zero gives the optimal clustering.

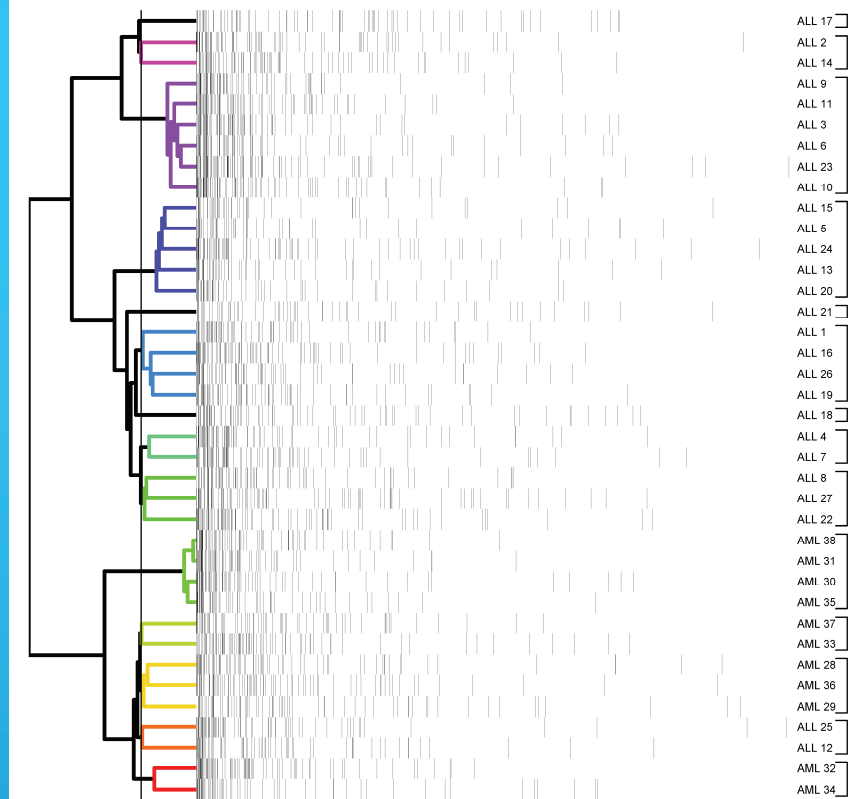
Metabolite Example

The goal of the study is to use Gas-Chromatography/Mass Spectrometry spectra for different plant phenotypes to classify forward genetic mutants of *Arabidopsis thaliana* (Messerli et. al. 2007). The metabolite data consist of 14 types of known, unknown and wild-type mutants, each comprising 43 metabolites. All types have four replicates; the other has three. The clustering method successfully separates wild types *ColWT*, *WsWT* and *RLDWT* from known mutants *pgm*, *isa2*, *sex1*, *sex4*, *mex1* and *dpe2*.



Microarray Example

We use the widely analysed leukemia data of Golub et. al. (1999), focusing on the 38 patients of the training set. The data are available through the *Bioconductor* package *golubEssets*. We selected 3051 of reasonably variant genes and normalized it to have zero median for each gene and unit overall variance. In the non-replicated case the proposed model is not identifiable. We use σ^2 as a tuning parameter and estimate the other hyperparameters using maximum marginal likelihood. The result for $\sigma^2=0.9$ reported below, successfully separates ALL from AML leukemia patients.



Apart from a one time optimisation needed to provide estimate of hyperparameters ($\mu, \sigma^2, \sigma_{\eta}^2, \sigma_{\theta}^2, p$), the number of computations required is $O(VT^3)$, linear in the number of variables V . The time needed for clustering the metabolite data was about 1/100 of a second and for the microarray data it was about 10 seconds on a PC with an Intel Centrino Duo processor, 1 GB RAM, and the Linux UBUNTU operating system.

References

- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) **Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring**, *Science* 286, 531–537.
- Messerli, G., Nia, V. P., Trevisan, M., Kolbe, A., Schauer, N., Geigenberger, P., Chen, J., Davison, A. C., Fernie, A. and Zeeman, S. C. (2007) **Rapid Classification of Phenotypic Mutants of Arabidopsis via Metabolite Fingerprinting**, *Plant Physiology* 143, 1484–1492.