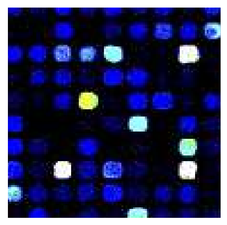


Bayesian Two-way Clustering for Gene Expression Data

University of BRISTOL

Introduction

Recent developments in genomics have led to the availability of a new type of data generated using microarray-based technologies. These data consist of measurements of expression levels of thousands of genes simultaneously, and can potentially provide information on the complex interplay of genes in an entire genome.

There are obvious attractions to a fully Bayesian hierarchical modelling approach to the uncovering of pattern in gene expression data. These include the possibility of seamless integration of the process of gene discovery with other inferential questions, avoidance of the hypothesis-testing paradigm for detection, full evaluation of uncertainty in inference, and the objectivity that comes from an analysis that is model-based, rather than defined only algorithmically.

However, there are equally obvious difficulties with attempting to do this. One is the challenge of modelling data exhibiting such heterogeneous variation (array-specific effects, gene-specific variances, etc.). There are also problems associated with MCMC computation for such models, exacerbated by the large numbers of variables, and issues in presenting aspects of high-dimensional posterior distributions in a visually informative way.

We describe our work so far in this endeavour, based on a two-way clustering model related to that of MacKay and Miskin (2001). The results are encouraging.

A Simple Model

MacKay and Miskin (2001) propose a model of the form

$$y_{gs} = \sum_{h=1}^{H} a_{sh} b_{gh} + \varepsilon_{gs},$$

where g =gene, y_{gs} is the data, $a_{sh} \in \mathbb{R}$ is a sample factor, $b_{gh} \in \mathbb{R}$ is a gene factor and ε_{gs} is noise. We think that this model is a little too general, so we simplify it.

We first consider a simple model for the single-sample case:

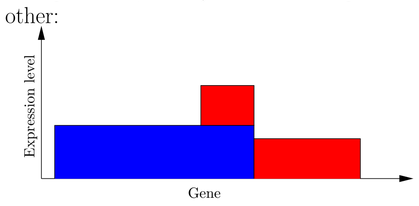
$$y_g = \sum_{h=1}^{H} z_{gh} b_h + \varepsilon_g,$$

where $z_{gh} \in \{0, 1\}$ is an allocation variable and $b_h \in \mathbb{R}$ is a constant level for layer h. This is very similar to the standard mixture model with constant variance:

$$y_g = \sum_{h=1}^{H} z_{gh} \mu_h + \varepsilon_g$$

except that the allocations for a mixture model are constrained to give $\sum_g z_{gh} \equiv 1$.

One of the key features of this model is that it allows layers to overlap, with a single gene being allocated to more than one layer. This is why we refer to them as "layers" rather than "components", as they are built on top of each other:

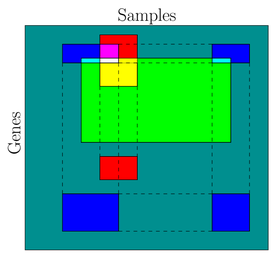


The full model

Having considered the single-sample case, we now extend the model to the multiple-sample case by adding per-sample allocations:

$$y_{gs} = \sum_{h=1}^{H} v_{sh} z_{gh} b_h + \varepsilon_{gs},$$

where s =sample and $v_{sh} \in \{0, 1\}$ is an allocation variable. This makes the layers constant-height rectangles as shown in the picture below.



Prior specification

We propose a fully Bayesian MCMC implementation of the models using the following priors:

$$\begin{aligned} z_{gh} &\sim \text{Bernoulli}\left(\tfrac{q}{H+1}\right) \\ v_{sh} &\sim \text{Bernoulli}(q) \\ b_h &\sim N(0, \tau^2) \\ H &\sim \text{Poisson}(\lambda) \\ \varepsilon &\sim N(0, \sigma^2). \end{aligned}$$

We also use a conjugate hyperprior on σ:

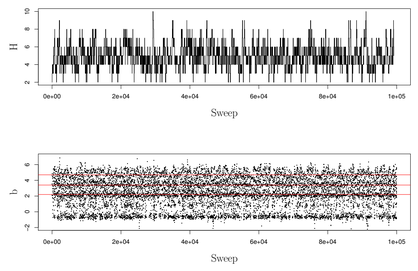
$$\sigma^{-2} \sim \Gamma(\alpha, \beta).$$

All of the other hyperparameters are held constant.

Implementation

We implemented an MCMC sampler of this model in `C++`. The sampler performs updates of z, v, b and σ using Gibbs sampler steps.

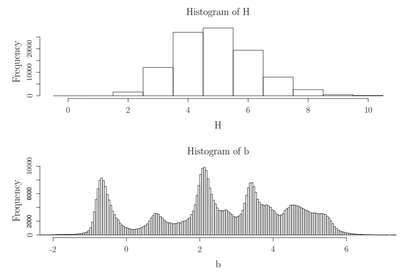
The number (and composition) of layers is updated using both birth-death and split-merge moves. There are two different split proposals: one to symmetric values of b_h, and a "bud" move, where one component retains the old value of b_h from the pre-split layer.



A Single Sample Example

We simulated data to see how successful the implementation was. The test data consisted of 500 genes with layers at 2.2, 3.4 and 4.7. There were also composite values at 5.6 (= 2.2+3.4), 6.9 (= 2.2+4.7), 8.1 (= 3.4+4.7) and 10.3 (= 2.2 + 3.4 + 4.7), as well as some values at 0.0. The s.d. of the noise was $\frac{1}{2}$.

Under this model the number of layers was over-estimated.

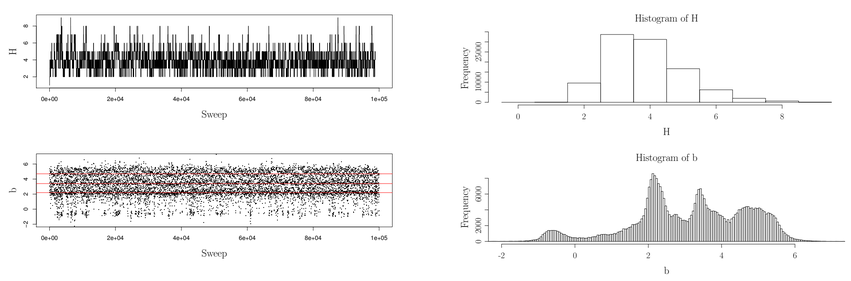


A repulsive prior(!)

We suspected that one of the reasons that the algorithm favoured an excessive number of layers was that there was little in the prior specification that favoured parsimony, i.e. two layers with values at 2.15 and 2.25 will almost certainly fit better than a single layer at 2.2.

In order to discourage this type of overfitting, we decided to change the prior on b and H to introduce some repulsion between the b-values using an area-interaction process (Baddeley and van Lieshout 1995).

After this modification, the sampler correctly found that 3 layers with values at about 2.2, 3.4 and 4.7 were sufficient to fit the data, though the position of the layer at 4.7 was rather uncertain.

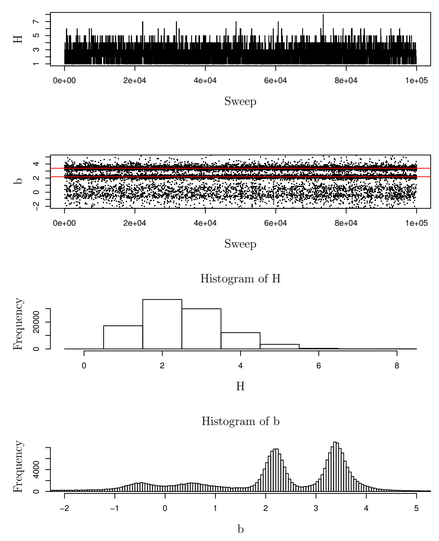


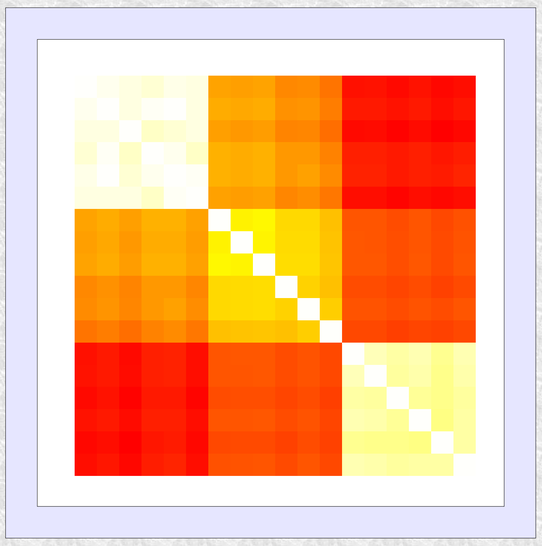
A toy example of the full model

We next simulated a dataset containing 10 genes for each of 5 samples with 2 layers at 2.2 and 3.4:

2.2	2.2	0	0	2.2
2.2	5.6	3.4	3.4	2.2
0	3.4	3.4	3.4	0
0	3.4	3.4	3.4	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	3.4	3.4	3.4	0
2.2	5.6	3.4	3.4	2.2
2.2	2.2	0	0	2.2

We added $N(0, (\frac{1}{2})^2)$ noise to this dataset as well. The sampler found the correct number and composition of the layers with high probability.





A real dataset

In order to prove our method with a real dataset, we looked at the human fibroblast data introduced by Lemon et al. (2002). This data set consists of 18 samples split into 3 categories: serum starved, serum stimulated and a 50:50 mix of starved/stimulated. We used the natural logarithm of Lemon et al.'s calculated LWF values as our measure of expression and subtracted gene and sample mean levels. We then selected the 100 most variable genes across all 18 samples and used this 18×100 array as the input to our sampler.

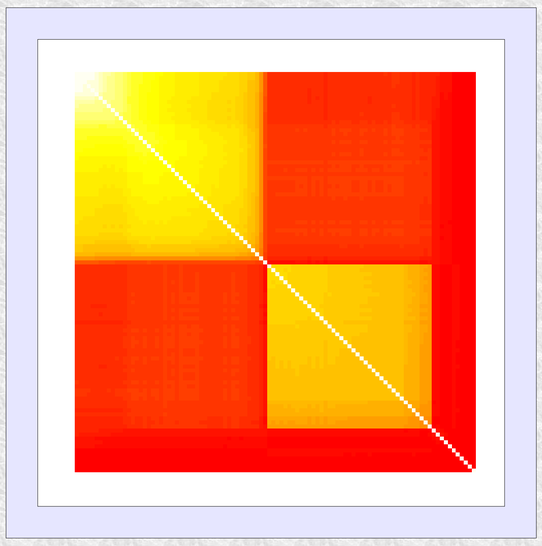
The picture on the left shows a similarity matrix of the posterior probability that a pair of samples are in a common layer. The three categories of sample are immediately obvious. The picture on the right shows the same statistic for genes, sorted using standard hierarchical clustering.

References

Baddeley, A. J. and M. N. M. van Lieshout (1995). Area-interaction point processes. *Annals of the Institute for Statistical Mathematics 47*, 601–619.

Lemon, W. J., J. J. T. Palatini, R. Krahe, and F. A. Wright (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. Accepted by *Bioinformatics*.

MacKay, D. J. C. and J. Miskin (2001). Latent variable models for gene expression data. Technical report, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE. United Kingdom.



Graeme Ambler, University of Bristol
School of Mathematics, University Walk, Bristol, UK, BS8 1TW.
Graeme.Ambler@bristol.ac.uk

Peter J. Green, University of Bristol
School of Mathematics, University Walk, Bristol, UK, BS8 1TW.
P.J.Green@bristol.ac.uk