

Bayesian Clustering and Model Exploration

Paul E. Anderson¹, Jim Q. Smith^{1*},
Kieron D. Edwards², and Andrew J. Millar²

¹ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.

² Institute of Molecular Plant Sciences, University of Edinburgh, EH9 3JH, UK.

Abstract. The arrival of longitudinal microarray data in biology demands the development of new types of clustering algorithms. Clustering is required over tens of thousands of time series (gene expression profiles) with perhaps only ten time points. Further, the experiments are designed to determine which genes exhibit a particular qualitative structure: we shall focus on circadian genes. An alternative to clustering over points in Euclidean space is thus needed. We modify a recent Bayesian clustering algorithm to address these issues. This adaptation employs the posterior distributions of the parameters in the Bayesian models. These were originally used to score cluster partitions. We propose their utility in categorising interesting clusters and then enlist this classification in a more effective and efficient search of the vast space of possible partitions. These methods are applicable to the clustering of any time series data.

1 Introduction

Clustering of gene expression profiles produced by microarray experiments are now not only performed on thousands of genes simultaneously, but also over a time course. Often the aim is to group together those genes that have a similar expression pattern so that existing genes can be used to identify unknown genes with an equivalent behaviour. See figure 1 for some typical profiles. In this way, the grand vision is to identify genes involved in the same regulatory network.

A recent Bayesian algorithm [1, 2] addresses this type of modified clustering problem, where time ordering must be respected, and appears to cluster on ramp changes in expression level over short time courses rather well. However, recent experiments have been performed over longer time courses that aim to uncover more complex profiles. Here we describe some recent and potential future developments of Bayesian clustering algorithms designed to search effectively and efficiently for genes whose behaviour is postulated to have a structured response, such as those involved in circadian (i.e. 24 hour) rhythms. The posterior parameters of fast Bayesian clustering algorithms can be used not only to score competing cluster partitions (as in [1]), but can also be used to *direct* the search algorithm itself. In particular, new staged clustering algorithms can be constructed to improve upon the basic search algorithms currently employed. In this paper, we present the development of one such algorithm.

* Author who will be presenting at the workshop. Correspondence should be sent to Anderson (p.e.anderson@warwick.ac.uk) and Smith (j.q.smith@warwick.ac.uk).

2 Bayesian Clustering Algorithms

First, we provide a brief summary of the model based approach expounded in [1], following the notation of [1, 2]. This Bayesian cluster methodology groups together those curves that appear to have been drawn from a joint distribution with parameters $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ under the linear model: $\mathbf{Y} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Here, \mathbf{Y} contains the gene expression profiles, \mathbf{B} is the design or basis function matrix that encodes the type of basis used for the clustering (linear spline, Fourier, wavelets, etc) and holds p basis functions, $\boldsymbol{\beta}$ is the parameter vector to be estimated and $\boldsymbol{\epsilon}$ is the usual Gaussian noise component, so that $\boldsymbol{\epsilon} \sim \text{N}(0, \sigma^2)$. One of the important features of the linear model structure is that it allows us to introduce time dependence. In [1] the chosen basis was a linear spline. For the application advanced here, a more appropriate choice is the Fourier basis, see below. Gene profiles in the same cluster are assumed to have originated from a common Gaussian process (this will become evident when we present the marginal likelihood) and will, within a given cluster c , have the same $\boldsymbol{\beta}_c$ and σ_c^2 . Note that $\boldsymbol{\beta}_c$, σ_c^2 and N_c (the number of genes in the cluster, c) will be different for each cluster, and that the dimensions of \mathbf{B} and \mathbf{y} will need to be adjusted accordingly to enable the calculations below to be carried out. (Specifically, \mathbf{B}_c will be an $N_c T \times p$ matrix and \mathbf{y}_c an $N_c T \times 1$ vector, assuming we have T time points.)

Under Bayesian clustering the score of a particular partition of clusters or *encoding function* is a linear function of the score of each of its cluster components. For each cluster c , the log marginal likelihood is

$$\lambda_c(\mathbf{y}) \equiv \log p_c(\mathbf{y}) = \log \left(\int \int p_c(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p_c(\boldsymbol{\beta}|\sigma^2) p_c(\sigma^2) d\boldsymbol{\beta} d\sigma^2 \right) \quad (1)$$

We thus need priors on $\boldsymbol{\beta}$ and σ^2 for each cluster in an encoding function. In this paper, for computational simplicity we follow [1] and choose these to be conjugate. For the Bayesian linear model above, this is the well-established normal inverse-gamma conjugate family, so that:

$$p_c(\boldsymbol{\beta}|\sigma^2) \sim \text{N}(\mathbf{m}, \sigma^2 \mathbf{V}) \quad \text{and} \quad p_c(\sigma^2) \sim \text{IGamma} \left(\frac{\alpha}{2}, \frac{\gamma}{2} \right) \quad (2)$$

\mathbf{V} is the prior covariance matrix (simply $v\mathbf{I}$ for some constant v) and \mathbf{m} is a vector.

Setting the prior mean $\mathbf{m} = 0$ to centre $\boldsymbol{\beta}$, standard results [2] tell us that the four parameters \mathbf{m} , \mathbf{V} , α and γ are updated in the following way:

$$p_c(\boldsymbol{\beta}|\mathbf{y}, \sigma^2) \sim \text{N}(\mathbf{m}^*, \sigma^2 \mathbf{V}^*) \quad \text{and} \quad p_c(\sigma^2|\mathbf{y}) \sim \text{IGamma} \left(\frac{NT + \alpha}{2}, \frac{d + \gamma}{2} \right) \quad (3)$$

with

$$\mathbf{V}^* = (\mathbf{B}'\mathbf{B} + \mathbf{V}^{-1})^{-1}, \quad \mathbf{m}^* = \mathbf{V}^* \mathbf{B}' \mathbf{y} \quad \text{and} \quad d = \mathbf{y}' (\mathbf{I} - \mathbf{B} \mathbf{V}^* \mathbf{B}') \mathbf{y} \quad (4)$$

' denotes the transpose and \mathbf{B} and \mathbf{V} are appropriately chosen, known matrices.

We can now use the marginal likelihood and prior predictive distribution $p_c(\mathbf{y})$ to score the clusters in any given encoding function. This can be specified analytically [1] for each cluster as a multivariate t-distribution:

$$p_c(\mathbf{y}) = \left(\frac{1}{\pi}\right)^{NT/2} \frac{\gamma^{\alpha/2} \Gamma\left(\frac{NT+\alpha}{2}\right) |\mathbf{V}^*|^{1/2}}{\Gamma\left(\frac{\alpha}{2}\right) |\mathbf{V}|^{1/2}} \frac{1}{(d+\gamma)^{(NT+\alpha)/2}} \quad (5)$$

$$= g(NT, \alpha, \gamma) |\mathbf{V}|^{-1/2} \frac{1}{|\mathbf{B}'\mathbf{B} + \mathbf{V}^{-1}|^{1/2} (d+\gamma)^{(NT+\alpha)/2}} \quad (6)$$

Let $C(\phi)$ denote the collection of clusters associated with the encoding function ϕ . Then under appropriate independence assumptions on the parameters of different clusters, the score $\Sigma(\phi)$ for any encoding function ϕ with clusters $c \in C(\phi)$ is given by

$$\Sigma(\phi) = \sum_{c \in C(\phi)} \log p_c(\mathbf{y}) + \Pi(\phi) \quad (7)$$

where $\Pi(\phi)$ is a known function of the prior distribution over encoding functions, typically chosen to depend only on the number of clusters in the partition defined by ϕ . Examples of good choices of $\Pi(\phi)$ are given in [1].

Useful encoders will have high scores. One pleasant feature of this class of models is that the difference between two encoding functions, identical outside a given set c , will depend only on their relative scores over this set of profiles. In particular, suppose two encoders differ only on a set $c = c_1 \cup c_2$ with $c_1 \cap c_2 = \emptyset$ where c is the level set of the encoding function ϕ^- and c_1, c_2 are two level sets on ϕ^+ . ϕ^+ contains the same clusters as ϕ^- except that cluster c in ϕ^- is separated into two clusters c_1 and c_2 in ϕ^+ . We say that two such encoders ϕ^+ and ϕ^- are *adjacent*. Thus,

$$\Sigma(\phi^+) - \Sigma(\phi^-) = \log p_{c_1}(\mathbf{y}_1) + \log p_{c_2}(\mathbf{y}_2) - \log p_c(\mathbf{y}) + \lambda(\phi^+, \phi^-) \quad (8)$$

where $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and $\lambda(\phi^+, \phi^-) = \Pi(\phi^+) - \Pi(\phi^-)$ is a priori often set to a fixed constant not depending on ϕ^+ or ϕ^- . Since for moderate sized clusters it is almost instantaneous to calculate the functions in the expression above, it is trivial to check whether merging or splitting two clusters to form a new encoder is more or less well supported by the data.

Currently, Heard *et al* [1] employ an agglomerative hierarchical technique (AHT) to search the space of encoders. Because this only modifies an encoder function to one adjacent to it in the sense above, even with large number of genes it is possible to quickly find encoders that cluster together genes with similar longitudinal profiles. In the AHT, we start with each of the N gene profiles in N separate clusters. We then form a sequence of new encoders by sequentially merging the two clusters that, when combined, increase the similarity measure by the largest amount, and repeat. The last encoder has one cluster with all N genes. Since we have calculated the marginal likelihood at each merger, we

can choose the one in this sequence that has the maximum score. This iterative optimisation of a marginal likelihood criterion is fast, but clearly leaves most of the encoder space unsearched. The next section describes an improved search algorithm which also moves between adjacent encoders. It uses various outputs of the prior to posterior analyses to evaluate the score of an encoder and thus helps guide the search.

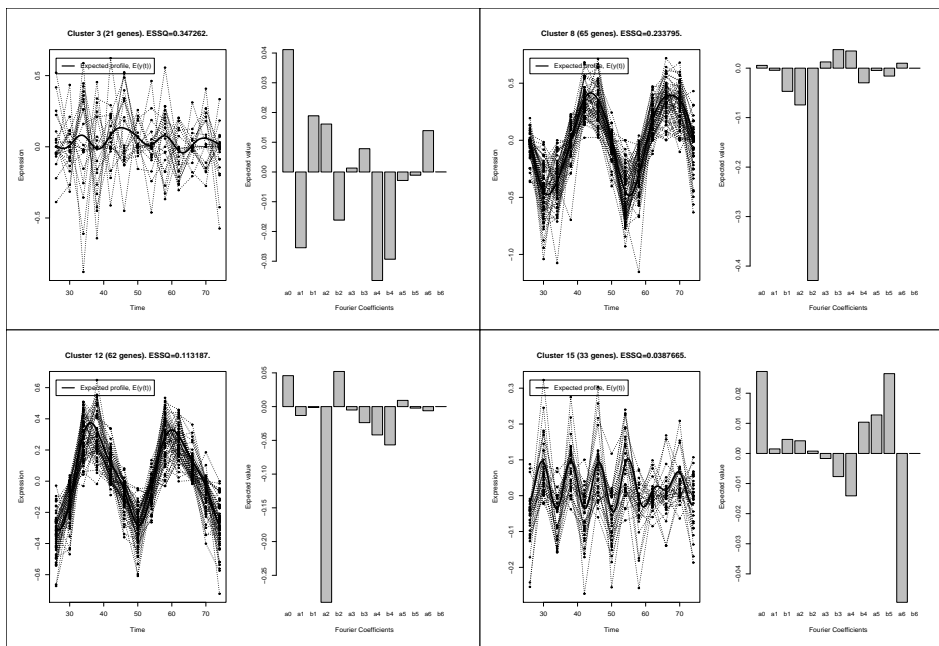


Fig. 1. Four typical clusters of gene expression profiles generated by agglomerative hierarchical clustering with a Fourier basis. The aim is to find circadian genes. Cluster 3 would be classified as junk, clusters 8 and 12 as interesting and cluster 15 as uninteresting. ESSQ is the estimated value of σ^2 . See the text for details.

3 Clocks and Clustering

Our analysis is of a recent microarray experiment on the plant *Arabidopsis thaliana* designed to detect genes whose expression levels, and hence functionality, may be connected with circadian rhythms. Fourier is the obvious choice of basis for this problem. Fortunately, the setting of hyperparameters for conjugate analyses with Fourier bases is well studied [3]. The time series is 48 hours long so we expect circadian genes to show two cycles. Microarrays are taken every four hours. A full analysis and exposition of this data, together with a discussion of its biological significance, will appear in a future biological publication.

However, our interests here are: to illustrate the nature of the clusters obtained from using the methodology of [1] with a Fourier basis; to study the useful

intermediate statistics that such an analysis routinely calculates; and to propose ways in which these statistics can then be used to improve the search for more useful encoder functions.

Figure 1 shows four typical clusters from the 31 produced by the encoder function with the best score using AHT. The gene expression profiles in each cluster are shown as dotted lines on the left hand side, with the posterior mean of σ^2 (a parameter describing the tightness of the cluster) given in each title. The expected profile of any cluster is linked to the expected value of the Fourier coefficients, here the means of the β vector, by

$$E(y(t)) = E(a_0) + \sum_{i=1}^6 \left(E(a_i) \cos\left(\frac{2\pi it}{48}\right) + E(b_i) \sin\left(\frac{2\pi it}{48}\right) \right) \quad (9)$$

The expected profile is the solid line on the profile plot, whilst a bar chart of the regression coefficients is displayed on the right hand side.

For a collection of genes to be interesting in this experiment, they must exhibit a strong circadian rhythm, see clusters 8 and 12 in figure 1. In our chosen basis this will be reflected in a larger estimated amplitude of the second harmonic, that is $(E(a_2)^2 + E(b_2)^2)^{\frac{1}{2}}$. Further, within this collection of clusters, we are interested in those that appear to respond most quickly to the stimulus: those whose first derivatives at zero are large. Again, this is reflected in a simple function of the expectation of the Fourier coefficients of that cluster.

In contrast, clusters like number 15 in figure 1 that are tight (have a smaller estimated value of σ^2) and do not exhibit a strong second harmonic are of little interest. There is no scientific merit in searching for encoders that further refine these clusters. Finally, our experience of working with the AHT in [1] has been that, because of the limitations in its search scope, junk clusters sometimes appear. These have diffuse profiles but are clustered together because their difference is explained by a large estimated value of σ^2 , see cluster 3 in figure 1. These clusters contain genes with profiles that would fit much better into other clusters. We are currently testing and coding cluster algorithms following the steps in the next section.

4 Customising the Search with Bayesian Techniques

1. Initialise the clustering and obtain the encoder ϕ with the highest score by applying the algorithm in [1] with a Fourier basis.
2. Identify the two subsets $I(\phi)$ and $J(\phi)$ with $I(\phi), J(\phi) \subseteq C(\phi)$ where $I(\phi)$ is the set of clusters of scientific interest (here clusters whose second harmonics are greater than an assigned threshold) and $J(\phi)$ is the set of junk clusters (those with a second harmonic below the threshold, but with a variance larger than a given threshold). For the reasons given above, our search algorithm does not waste time trying to refine clusterings involving genes not in $I(\phi)$ or $J(\phi)$.
3. Search iteratively for a better encoder than the current best encoder ϕ by sieving and sorting until the algorithm converges.

- To *sieve*: construct a new encoder ϕ_J^- whose clusters $C(\phi_J^-)$ consist of $c \in J(\phi)$ together with a singleton cluster for each gene in $J(\phi)$. Using ϕ_J^- as a starting point, use AHT to find the encoder ϕ_J with the highest score $\Sigma(\phi_J)$ and replace ϕ by ϕ_J if $\Sigma(\phi_J) > \Sigma(\phi)$. If ϕ_J is adopted then remove any tight uninteresting clusters to form a smaller junk set $J(\phi_J)$.
- To *sort*: for a chosen $c \in I(\phi)$, partition $c = c_1 \cup c_2$ to obtain a new encoder ϕ_S^- whose clusters agree with ϕ except on c where it takes two levels: labelling genes as in c_1 or c_2 . If $\Sigma(I(\phi_S^-)) > \Sigma(I(\phi))$, replace ϕ by ϕ_S^- . See [4] for more details on such a step.

Repeat this procedure a fixed number of times or until at least one change is made. Use AHT with the new encoder as a starting point and identify the encoder ϕ_S with the highest score. Currently, both $c \in I(\phi)$ and c_1, c_2 are chosen at random.

Note that this is fully Bayesian in the sense that we are still calculating posterior predictive scores and optimising over them. However, we are searching many more cluster configurations than under greedy search. Also, estimates of the parameters of the best encoder ϕ are being used to guide our search for a better encoder by identifying $I(\phi)$ and $J(\phi)$.

5 Discussion

This Bayesian clustering method is still in its early stages but initial results are promising. It offers the possibility of clustering longitudinal data whilst respecting the temporal order, something not possible with Euclidean clustering. Our proposed adjusted method appears to identify tighter clusters than those based simply on agglomerative hierarchical clustering and is more robust to augmentation of the time series (here by inclusion or deletion of subsets of gene profiles). Our results on these and other developments will be reported in a future paper.

We thank Nick Heard for useful discussions. Paul E. Anderson, Jim Q. Smith and Andrew J. Millar are part of the Interdisciplinary Program for Cellular Regulation (IPCR) based at the University of Warwick.

References

1. Heard, N.A., Holmes, C.C., Stephens, D.A.: A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. To appear in the Journal of the American Statistical Association (2005)
2. Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M.: Bayesian Methods for Nonlinear Classification and Regression. Wiley Series in Probability and Statistics. John Wiley and Sons (2002)
3. Campodónico, S., Singpurwalla, N.D.: The Signature as a Covariate in Reliability and Biometry. In Freeman, P.R., Smith, A.F.M., eds.: Aspects of Uncertainty — A Tribute to D. V. Lindley. John Wiley and Sons (1994) 119–147
4. Chipman, H., George, E., McCullough, R.: Bayesian CART Model Search. Journal of the American Statistical Association **93** (1998) 935–960