Bayesian Analysis (2009) 4, Number 3, pp. 539–572

Efficient Utility-based Clustering over High Dimensional Partition Spaces

Silvia Liverani*, Paul E. Anderson†, Kieron D. Edwards‡,
Andrew J. Millar§ and Jim Q. Smith¶

Abstract. Because of the huge number of partitions of even a moderately sized dataset, even when Bayes factors have a closed form, in model-based clustering a comprehensive search for the highest scoring (MAP) partition is usually impossible. However, when each cluster in a partition has a signature and it is known that some signatures are of scientific interest whilst others are not, it is possible, within a Bayesian framework, to develop search algorithms which are guided by these cluster signatures. Such algorithms can be expected to find better partitions more quickly. In this paper we develop a framework within which these ideas can be formalized. We then briefly illustrate the efficacy of the proposed guided search on a microarray time course data set where the clustering objective is to identify clusters of genes with different types of circadian expression profiles.

1 Introduction

Many Bayesian model selection procedures are based on the posterior probability distribution over models. Two very common methods are MAP selection, where the most a posteriori probable model is selected (Heard et al. 2006), and model mixing, where posterior probability distributions are calculated over the most promising candidate models and the results then mixed over these models (Fraley and Raftery 1998). Here, for simplicity we will focus on the former. In either case, a full exploration of the partition space is not possible when, as in our case, the number of elements in a cluster is in the order of tens of thousands, even when using fast conjugate modelling. The number of partitions of a set of n elements grows quickly with n. For example, there are 5.1×10^{13} ways to partition 20 elements.

he problem with MAP search is that it scores two partitions differently even when they agree on all the expressions of scientific interest, even if the cluster objects are prefiltered. This therefore means that the vast proportion of the search is used comparing the efficacy of one partition with another when their difference is of no scientific consequence. For example, in our running example, but using MAP search, an algorithm will spend a vast proportion of its time trying to cluster obviously non-circadian genes in the best possible way: clearly a futile waste of time in the context of this experiment.

*Department of Statistics, University of Warwick, UK, mailto:S.Liverani@warwick.ac.uk
†Department of Statistics, University of Warwick, UK, mailto:p.e.anderson@warwick.ac.uk
‡Advanced Technologies Limited, Cambridge, UK
§Centre for Systems Biology at Edinburgh, University of Edinburgh, UK, mailto:Andrew.Millar@ed.ac.uk
¶Department of Statistics, University of Warwick, UK, mailto:J.Q.Smith@warwick.ac.uk

 DOI:10.1214/09-BA420

The objective of this paper is to find search algorithms which share the possibility of fast local search with MAP algorithms whilst avoiding the type of inefficacy discussed above.

In our paper we assume that each cluster has a signature defining how scientifically interesting each cluster is. This context is quite common and in our case it was motivated by the need to cluster data from time course microarray data. Many algorithms have been used to cluster such datasets, in both the classical (Banfield and Raftery 1993; Fraley and Raftery 1998; Ben-Dor et al. 1999; Yeung et al. 2001; Ramoni et al. 2002; Luan and Li 2003) and Bayesian framework (Wakefield et al. 2003; Heard et al. 2006; Zhou et al. 2006). For experiments with multiple microarray experiments (i.e. with different doses, different treatments or followed up over time), Bayesian algorithms are very versatile, for example to model time dependence between microarrays and to enable incorporation of pertinent scientific information. See e.g. Heard et al. (2006) and Zhou et al. (2006).

In our running example the scientists were only interested in discovering those genes whose expression profiles over two days exhibited circadian rhythms: other expression profiles were irrelevant. Because of the enormous size of the search space, for the sake of efficiency, it looked promising to try to customize the search algorithms so that they reflect the scientific inquiry by focusing an algorithm to refine only clusters containing potentially interesting genes and not to waste time refining parts of a partition of no interest to the scientist. The question we answer in this paper is: can this sort of procedure be formalized within a Bayesian framework?

MAP selection has a utility based justification (Bernardo and Smith 1994) and Bayesian selection techniques with specific priors have been limited to different classes of score functions (Denison et al. 2002). However there is a more pertinent literature - albeit specifically for Bayesian Networks, e.g. Tatman and Shachter (1990) springing from a subclass of multiattribute utility functions. This describes how, when the decision maker's utility function is separable, then, with the appropriate structure of prior, the search for an optimal decision can be localized, facilitating fast optimization. Using a utility led approach we will demonstrate how a modification of this algorithm also allows us to focus search on parts of the parameter space of importance to the scientist.

The paper is organized as follows. In Section 2 we present the class of conjugate Gaussian regression models as in Smith et al. (2008): one of the types of model to which our methodology can apply. In Section 3 we introduce a formal framing of this genre of clustering problems in terms of multiattribute decision theory and discover a set of assumptions that will lead us to formally explain only parts of the underlying partition space. In Section 4 we show that if the product utility function is used, local search algorithms, widely used for conventional model exploration, are equally valid within this general framework. This means that the new utility based method is easy to implement. In Section 5 we briefly illustrate these methods through two examples. These concern a recent microrray experiment on the plant model organism *Arabidopsis thaliana*, designed to detect genes whose expression levels, and hence functionality, might be connected with circadian rhythms. The examples describe how our utility methods

can be quickly applied to a very large dataset: here over 22,000 13-dimensional profiles were clustered.

2 A Clustering for Time Course Data

For the sake of simplicity, in this paper we illustrate our utility based approach in conjuction with a conjugate model developed by Heard et al. (2006) which we found particularly appealing. It is fast, flexible and it includes all the features of our data in the Gaussian conjugate Bayesian regression model. We next summarise some of the analysis in Smith et al. (2008) which will be intrinsic to the later development of this paper. Let $\mathcal{D} = (Y_1, \ldots, Y_n)$ and $Y = \text{vec}(\mathcal{D})$ satisfy

$$Y = X\boldsymbol{\beta} + \varepsilon$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)' \in \mathbf{R}^p$ and $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$ is a vector of independent error terms with $\sigma^2 > 0$. Note that $Y_i \in \mathbf{R}^{\mathcal{T}}$ for $i = 1, \ldots, n$. The posterior Normal Inverse Gamma joint density of the parameters $(\boldsymbol{\beta}, \sigma^2)$ denoted by $NIG(\mathbf{0}, V, a, b)$, is given by

$$p(\boldsymbol{\beta}, \sigma^2|y) \propto (\sigma^2)^{-(a^*+p/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - m^*)'(V^*)^{-1}(\boldsymbol{\beta} - m^*) + 2b^*\right]\right\}$$

with

$$\begin{array}{rclcrcl} m^* & = & (V^{-1} + X'X)^{-1}X'Y & & a^* & = & a + rn/2 \\ V^* & = & (V^{-1} + X'X)^{-1} & & b^* & = & b + \gamma/2 \\ \gamma & = & \{Y'Y - (m^*)'(V^*)^{-1}m^*\} & & & & \end{array}$$

where $a, b > 0$ and V is a positive definite matrix. Throughout this paper we assume that $X = \mathbf{1}_n \otimes \mathcal{B}$, where $\mathcal{B}$ is a known matrix, and that $X'X = n\mathcal{B}'\mathcal{B}$ is full rank. The design or basis function matrix $\mathcal{B}$ encodes the type of basis used for the clustering: linear splines in Heard et al. (2006), wavelets in Ray and Mallick (2006) or Fourier in Edwards et al. (2006): the last one is the most appropriate choice in the context of a study of daily rhythms of genes we study in our illustrative examples.

The Bayes factor can be calculated from its marginal likelihood $L(y)$ as in Denison et al. (2002) and O'Hagan and Forster (2004). Thus

$$L(y) = \left(\frac{1}{\pi}\right)^{n\mathcal{T}/2} \frac{b^a}{(b^*)^{a^*}} \frac{|V^*|^{1/2}}{|V|^{1/2}} \frac{\Gamma(a^*)}{\Gamma(a)}$$

Let C denote a partition belonging to the space of partitions $\mathcal{C}$, on a space Ω of cardinality n, and c a cluster of such a partition. Heard et al. (2006) assume that each gene profile is exchangeable within the cluster to which it belongs and consider the Normal Inverse-Gamma conjugate Bayesian linear regression model, which takes the form

$$Y^{(c)} = X^{(c)}\boldsymbol{\beta}^{(c)} + \varepsilon^{(c)} \tag{1}$$

for cluster c, where $\boldsymbol{\beta}^{(c)} = (\boldsymbol{\beta}_1^{(c)}, \ldots, \boldsymbol{\beta}_p^{(c)})$ is the vector of parameters with $p \leq \mathcal{T}$, $X^{(c)}$ is the design matrix of size $n(c)\mathcal{T} \times p$, $\varepsilon^{(c)} \sim N(0, \sigma_c^2 I_{\mathcal{T}n(c)})$ where $n(c)$ is number of

genes in cluster c and $I_{\mathcal{T}n(c)}$ is the identity matrix of size $\mathcal{T}n(c) \times \mathcal{T}n(c)$. A partition C of the genes divides into N clusters of sizes $\{n(c_1), \ldots, n(c_N)\}$, with $n = \sum_{i=1}^{N} n(c_i)$.

Although other priors can be used in this context, Smith et al. (2008) recommended the use of coherence priors over the partition space. Under these cohesion priors both the prior and posterior probability $\pi(C)$, where the generating partition is C, has the form

$$\pi(C|y) = A \prod_{c \in C} \pi(c|y) \tag{2}$$

where A is a constant ensuring the probabilities of different possible partitions all sum to one.

Assuming the parameters of different clusters are independent, because the likelihood separates, it is straightforward to check (Smith et al. 2008) that the log marginal likelihood score $\Sigma(C)$ for any partition C with clusters $c \in C$ is given by

$$\Sigma(C) = \sum_{c \in C} \log p_c(y) + \log \pi(C) \tag{3}$$

where $\log \pi(C)$ is given in (2).

An essential property of the search for MAP models - dramatically increasing the efficiency of the partition search - is that with the right family of priors the search is *local*. That is, if $\widehat{C}^+$ and $\widehat{C}^-$ differ only in the sense that the cluster $\widehat{c}^+ \in \widehat{C}^+$ is split into two clusters $\widehat{c_1}^-, \widehat{c_2}^- \in \widehat{C}^-$ then the log marginal likelihood score is a linear function only of the posterior cluster probabilities on $\widehat{c}^+, \widehat{c_1}^-$ and $\widehat{c_2}^-$. We show in Section 4 that this local property is preserved when we use our utility based clustering method provided a product utility search is employed.

The simplest search method using local search is agglomerative hierarchical clustering (AHC). It starts with all the genes in separate clusters, our original C_0, and evaluates the score of this partition. Each cluster is then compared with all the other clusters and the two clusters which increase the log likelihood in (3) by the most are combined to produce a new partition C_1. We now substitute C_1 for C_0 and repeat this procedure to obtain a partition C_2. We continue in this way until we have evaluated the logmarginal score $\Sigma(C_i)$ for each partition $\{C_i, 1 \leq i \leq n\}$. We then choose the partition which maximizes the score $\Sigma(C_i)$.

A drawback of this method and ones like it is that the set of searched partitions is an extremely small subset of the set of all partitions. Moreover, no regard is taken by the algorithm of whether there is any scientific inferential merit in combining two clusters together. In our context an automatic search algorithm like AHC will spend the vast majority of its time examining the efficacy of combining two non-circadian gene clusters, an activity quite worthless from the scientific perspective. The motivation of this paper is to try to find formal and efficient ways of addressing this obvious inadequacy of simple deterministic search.

3 Utility over partitions

3.1 A useful class of utilities

Our idea in this paper is to use a utility function expressing the nature of the scientific interest to guide the search for the partition focusing on finding the partition with the highest posterior expected utility.

Let us generalize the notation introduced earlier for our running example. Let $\boldsymbol{\theta}_c$ be the vector of parameters associated with a cluster c. In our running example $\boldsymbol{\theta}_c$ is the vector of regression coefficients $\boldsymbol{\beta}_c$ and the variance term σ_c^2 . Let $\boldsymbol{\theta}(C) = \{\boldsymbol{\theta}_c : c \in C\}$ denote the vector of parameters associated with a given partition. Recall that under the usual model assumptions - both a priori and a posteriori - the density $\pi_c(\boldsymbol{\theta}_c)$ of $\boldsymbol{\theta}_c$ depends on the cluster index c but not on the partition C and that the vectors $\{\boldsymbol{\theta}_c : c \in C\}$ are mutually independent of each other. Using this more general notation, it follows that the density $\pi(\boldsymbol{\theta}(C)|C, y)$ can be written in the form

$$\pi(\boldsymbol{\theta}(C)|C, y) = \prod_{c \in C} \pi_c(\boldsymbol{\theta}_c|C, y) = \prod_{c \in C} \pi_c(\boldsymbol{\theta}_c|y) \tag{4}$$

The most complex family $\mathbb{U}$ of utility functions over many attributes in current use consists of utility functions $U(\widehat{C}|C, \boldsymbol{\theta}(C))$ which exhibit mutually utility independent attributes (Keeney and Raiffa 1976; Keeney and von Winterfeldt 2007; French and Rios Insua 2000). In our context, when each attribute is the expression profile of each gene $i \in \Omega$, by definition these utilities have the functional form

$$U\left(\widehat{C}|C, \boldsymbol{\theta}(C)\right) + 1 = \prod_{i \in \Omega} (1 + \kappa_i u_i(\widehat{c}|c, \boldsymbol{\theta}_c)) \tag{5}$$

where the conditional utility $u_i(\widehat{c}|c, \boldsymbol{\theta}_c)$ is the utility score of gene i when placed in cluster $\widehat{c}$ when in the generating partition C gene i lies in cluster $c \in C$.

Recall here that the relative magnitude of κ_i to κ_j reflects the importance the scientist places on gene i relative to gene j, and that as $\max \kappa_i \to 0$ this utility function tends to a linear one, whilst as $\min \kappa_i \to \infty$ we only score partitions which succeed in classifying *all* genes partially well. We now identify a subclass $\mathbb{V} \subseteq \mathbb{U}$ that on the one hand can plausibly embody the preference structure of a typical biologist investigating gene profiles and on the other provides a framework for more focused search algorithms over the partition space.

Thus suppose the scientist is prepared to state whether each given gene $i \in \Omega$ is potentially interesting - henceforth written $i \in I$ - or uninteresting - denoted here by $i \in \overline{I}$. When $U \in \mathbb{U}$, the implication of the above is that the scientist should set $\kappa_i = 0$ whenever $i \in \overline{I}$. Note that sometimes it will be appropriate to set $I = \Omega$. Let n_I denote the number of potentially interesting genes.

Definition 3.1. *Say a partition C of Ω is* $I-$simple *if all of its clusters c either have the property $c \cap I = c$ or $c \cap I = \emptyset$. Denote the set of all $I-$simple partitions by $\mathcal{S}(I)$.*

A partition C is $I-$simple if and only if I can be expressed in the form

$$I = \bigcup_{c \in C(I)} c \tag{6}$$

where $C(I)$ is a subset of the clusters c of C such that $c \cap I = c$. Obviously all partitions are $I-$simple when $I = \Omega$. Clearly for any partition $\widehat{C}_1$ there is a partition $\widehat{C}_2 \in \mathcal{S}(I)$ such that $U(\widehat{C}_1|C, \boldsymbol{\theta}(C)) = U(\widehat{C}_2|C, \boldsymbol{\theta}(C))$. We henceforth restrict our search to the partitions $\mathcal{C}$ that belong to $\mathcal{S}(I)$.

Let π_I denote the probability under the mass function (2) that the generating partition $C \in \mathcal{S}(I)$. Then, if the scientist believes that $\pi_I = 1$, for any cluster c that does not satisfy $c \cap I = c$ or $c \cap I = \emptyset$, $\pi(c) = \pi(c|y) = 0$. A scientist making this assumption a priori believes that with probability one the generating partition will contain only clusters that inherit the label of being unambiguously interesting (i.e. containing only interesting genes) or unambiguously uninteresting (i.e. containing only interesting genes). This is a substantive but often plausible assumption. It embodies the belief that the definition of the term interesting is consistent with the underlying generating partition. If the scientist were not to hold this belief then it would bring into question whether a partition model should be used at all in the decision analysis. Note that this assumption simplifies the analysis because it allows the focus of the problem to switch from the individual units to the more coarse clusters of a partition.

It follows that we can write

$$\pi(C|y) = \pi(C|y, C \in \mathcal{S}(I)) = \pi_1(C(\overline{I})|y)\pi_2(C(I)|y) \tag{7}$$

where $C(\overline{I})$ is a partition of $\overline{I}$ and $C(I)$ is a partition of I and

$$\begin{aligned} \pi_1(C(\overline{I})|y) &= A_1 \prod_{c \notin C(I)} \pi(c|y) \\ \pi_2(C(I)|y) &= A_2 \prod_{c \in C(I)} \pi(c|y) \end{aligned} \tag{8}$$

where $A_1$and A_2 are proportionality constants ensuring $\pi_1(C(\overline{I})|y)$ and $\pi_2(C(I)|y)$ are probability mass functions. So in particular any function of C depending only on the configuration of clusters in the partition $C(I)$ of the interesting genes I and not those in $C(\overline{I})$ of $\overline{I}$ will be independent of $C(\overline{I})$.

Say that preferences are *cluster critical* if whenever $i \in \widehat{c} \neq c$ for all values of $\boldsymbol{\theta}_c$

$$u_i(\widehat{c}|c, \boldsymbol{\theta}_c) = 0 \tag{9}$$

A biologist's preferences will be consistent with this if for any gene i in the cluster $\widehat{c} \in \widehat{C}(I)$ of interesting genes to contribute to the utility score, it is necessary for i to be classified correctly so that $\widehat{c} = c$. When the conditional utilities are cluster critical write

$$w_i(\widehat{c}|\boldsymbol{\theta}_{\widehat{c}}) \triangleq u_i(\widehat{c}|\widehat{c}, \boldsymbol{\theta}_{\widehat{c}}) \tag{10}$$

In this paper we will also assume that the scientist's preferences over interesting genes within the same cluster are exchangeable. Thus assume that genes in I are *cluster exchangeable* meaning that

$$\kappa_i = \begin{cases} \kappa\phi(\widehat{c}) & \text{when } i \in I \cap \widehat{c} \\ 0 & \text{when } i \in \overline{I} \end{cases} \tag{11}$$

and that the genes in I that are cluster critical are *utility exchangeable* meaning that whenever $i, j \in \widehat{c}$

$$w_i(\widehat{c}|\boldsymbol{\theta}_{\widehat{c}}) = w_j(\widehat{c}|\boldsymbol{\theta}_{\widehat{c}}) \triangleq w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}) \tag{12}$$

where the functions of the conditional utilities $0 \leq w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}) \leq 1$ reflect how highly the gene $i \in \widehat{c}$ scores when i really lies in the cluster $\widehat{c}$ - with associated parameters θ_c - of the generating partition C. Note that a least preferable estimate $\widehat{c}$ of c has $w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}) = 0$ and the most preferable $w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}) = 1$.

In Section 5 we use the functions $w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}})$ to preferentially weight the score of some genes in a potentially interesting cluster in terms of the values of the parameters $\boldsymbol{\theta}_{\widehat{c}}$: for example those with high expression and/or parameter values that suggest a clear diurnal pattern that would be associated with circadian regulatory genes. In particular, in Section 5 we will aproximate this utility function by using a measure of the circadianity of genes over time. Further discussion on $w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}})$ is given in Section 4.2.

Definition 3.2. *Say a utility function $U \in \mathbb{U}$ is* amenable *if it is cluster critical and cluster and utility exchangeable - i.e. if the three conditions (9), (11) and (12) all hold. Denote the class of amenable utility functions by $\mathbb{V}$.*

Henceforth assume $U \in \mathbb{V}$. Then by definition, from (5) for decision $\widehat{C} \in \mathcal{S}(I)$

$$U\left(\widehat{C}|C, \boldsymbol{\theta}(C)\right) + 1 = \prod_{i \in I}(1 + \kappa_i u_i(\widehat{c}|c, \boldsymbol{\theta}_c)) \tag{13}$$

which by cluster criticality can be written

$$U\left(\widehat{C}|C, \boldsymbol{\theta}(C)\right) + 1 = \prod_{i \in J(\widehat{C})}(1 + \kappa_i w_i(\widehat{c}|\boldsymbol{\theta}_{\widehat{c}})) \tag{14}$$

where $J(\widehat{C})$ is the set of genes correctly classified by $\widehat{C}$ i.e.

$$J(\widehat{C}) = \{i : i \in \widehat{c} = c\} \tag{15}$$

By cluster and utility exchangeability this now reduces to the form

$$U\left(\widehat{C}|C, \boldsymbol{\theta}(C)\right) = \prod_{\widehat{c} \in \widehat{C}(C,I)}(1 + \kappa\phi(\widehat{c})w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}))^{n(\widehat{c})} - 1 \tag{16}$$

where $\widehat{C}(C, I)$ is the set of clusters correctly classified in $\widehat{C}(I)$.

For reasons that will become apparent later we will call a scientist's preference weights $\phi(\widehat{c})$ *balanced* if $\phi(\widehat{c}) = n(\widehat{c})^{n(\widehat{c})^{-1}}$.

3.2 Marginal Search

When $U \in \mathbb{V}$, from (4) and (16) the expected utility $\overline{U}(\widehat{C}|C)$ of choosing the partition $\widehat{C} \in \mathcal{S}(I)$, when the true generating partition is C, is given by

$$\begin{aligned} \overline{U}(\widehat{C}|C) &= \int U(\widehat{C}|C, \boldsymbol{\theta}(C))\pi\left(\boldsymbol{\theta}(C)|C, y\right) d\boldsymbol{\theta}(C) \\ &= \prod_{\widehat{c}\in\widehat{C}(C,I)} \overline{u}(\widehat{c}) - 1 \end{aligned} \tag{17}$$

where for each $\widehat{c} \in \widehat{C}(I)$

$$\overline{u}(\widehat{c}) = \int (1 + \kappa\phi(\widehat{c})w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}}))^{n(\widehat{c})}\pi_{\widehat{c}}\left(\boldsymbol{\theta}_{\widehat{c}}|y\right) d\boldsymbol{\theta}_{\widehat{c}} \tag{18}$$

Thus when the generating cluster is known, one plus the score of a simple partition is the product over scores obtained from each correct potentially interesting cluster.

Recall that $\widehat{C}(C, I) \subseteq C(I)$ is the set of clusters correctly classified in $\widehat{C}(I)$ and $C(I)$ is a partition of I. Because by definition $\widehat{C}(C, I) \subseteq C(I)$ note the score $\overline{U}(\widehat{C}|C)$ depends on C only through $C(I)$ and is independent of $C(\overline{I})$ because I and $\overline{I}$ are disjoint sets. By the comments after (8) if the scientist a priori believes $\pi_I = 1$ then the expectation $\overline{U}(\widehat{C})$ of $\overline{U}(\widehat{C}|C)$ over C - the score we assign to $\widehat{C}$ - only depends on our choice of $\widehat{C}(I)$. In this scenario, investigating splits and combinations of clusters in $C(\overline{I})$ is pointless since such moves cannot improve the score. Thus when $U \in \mathbb{V}$ and $\pi_I = 1$ there is no loss in restricting our moves between partitions $\widehat{C}^+$ and $\widehat{C}^-$ whose differential clusters $\widehat{c}^+ \in \widehat{C}^+$ and $\widehat{c_1}^-, \widehat{c_2}^- \in \widehat{C}^-$ lie in I.

Under the assumptions above we can therefore, without loss, simply search the partition space over the space I. However, in general, standard local search algorithms cannot be used for expected utility maximization because the local properties of this score function are lost. Nevertheless, in the next section we prove that the product utility function, which is a limit of the usual class of utility independent utilities, *does* retain this important property.

4 Properties of the product utility

4.1 Product Utilities and Local Moves

The product utility function is closely linked to the family $\mathbb{V}$ and also admits the simple evaluation of relative scores under local search.

Definition 4.1. *The* product utility function $U_I(\widehat{C}|C, \boldsymbol{\theta}(C))$ *on a set* $I \subseteq \Omega$ *has the form*

$$U_I(\widehat{C}|C, \boldsymbol{\theta}(C)) = \prod_{c\in C(I)} \phi(\widehat{c})^{n(\widehat{c})} u_{\widehat{c}}(\widehat{c}|c, \boldsymbol{\theta}_c)^{n(\widehat{c})} \tag{19}$$

where the conditional utilities $u_{\widehat{c}}(\widehat{c}|c, \boldsymbol{\theta}_c)$ are cluster critical. Denote the set of product utility functions on I by $\mathbb{V}_I$.

Note that in the notation developed above we can write a product utility function in the simplified form

$$U_I\left(\widehat{C}|C, \boldsymbol{\theta}(C)\right) = \begin{cases} \prod_{\widehat{c}\in\widehat{C}(I)} \left\{\phi(\widehat{c}) w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}})\right\}^{n(\widehat{c})} & \text{when } \widehat{C}(I) = C(I) \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $w_{\widehat{c}}(\boldsymbol{\theta}_{\widehat{c}})$ is defined in (10).

The product utility function $U_I \in \mathbb{V}_I$ is a limit of a utility function $U \in \mathbb{V}$ in the following sense. For any partition $C \in \mathcal{S}(I)$ write

$$\overline{U}_I(\widehat{C}|C) = \prod_{\widehat{c}\in\widehat{C}(I)} v(\widehat{c}) \tag{21}$$

where for each $\widehat{c} \in \widehat{C}(I)$

$$v(\widehat{c}) = \phi(\widehat{c})^{n(\widehat{c})} \int u_{\widehat{c}}(\widehat{c}|c, \boldsymbol{\theta}_c)^{n(\widehat{c})} \pi_c\left(\boldsymbol{\theta}_c|y\right) d\boldsymbol{\theta}_c \tag{22}$$

Recalling that n_I is the number of genes in interesting clusters, using the notation above we see that as $\kappa \to \infty$, and holding weights so that $\min\{\phi(\widehat{c}) : \widehat{c} \subseteq I\} \geq M$

$$\begin{aligned} \kappa^{-n_I}\left\{\overline{U}\left(\widehat{C}|C\right) + 1\right\} &= \prod_{\widehat{c}\in\widehat{C}(I)} \int \left(\kappa^{-1} + \phi(\widehat{c}) u_{\widehat{c}}\left(\widehat{c}|\widehat{c}, \boldsymbol{\theta}_c\right)\right)^{n(\widehat{c})} \pi_c\left(\boldsymbol{\theta}_c|y\right) d\boldsymbol{\theta}_{\widehat{c}} \\ &= \overline{U}_I(\widehat{C}|C) + 0(\kappa^{-1}) \end{aligned} \tag{23}$$

provided $u_{\widehat{c}}(\widehat{c}|\widehat{c}, \boldsymbol{\theta}_c) \geq \varepsilon > 0$ for all $c \in C$. So a linear transformation of the expected utility score of a sequence of $U \in \mathbb{V}$ tends to that of a corresponding product utility score as the criterion weight on all the potentially interesting genes becomes large.

In addition to inheriting the interpretability of its parameters from $\mathbb{V}$ the $U_I \in \mathbb{V}_I$ also exhibits the property that its scoring is local. Because $U_I(\widehat{C}|C, \boldsymbol{\theta}(C)) = 0$, whenever $\widehat{C} \neq C$ letting

$$\overline{U}_I(\widehat{C}) \triangleq \mathbb{E}\{\overline{U}_I(\widehat{C}|C)\} \tag{24}$$

- the expected value of $\overline{U}_I(\widehat{C}|C)$ over the possible generating partitions - we note that, since $\widehat{C} \in \mathcal{S}(I)$, with any cohesion prior (2) on the partitions

$$\begin{aligned} \overline{U}_I(\widehat{C}) &= \overline{U}_I(\widehat{C}|C)\pi(\widehat{C} = C, C \in \mathcal{S}(I)|y) \\ &= \overline{U}_I(\widehat{C}|C)\pi(\widehat{C} = C|y, C \in \mathcal{S}(I))\pi_I|y \\ &= \prod_{\widehat{c}\in\widehat{C}(I)} v(\widehat{c})\pi_2(\widehat{C}(I)|y)\pi_I|y \end{aligned}$$

So in particular on comparing the adjacent partitions $\widehat{C}^+$ and $\widehat{C}^- \in \mathcal{S}(I)$, $\overline{U}_I(\widehat{C}^+) \geq \overline{U}_I(\widehat{C}^-)$ if and only if

$$\log v(\widehat{c}^+) + \log \pi(\widehat{c}^+|y) \geq \log v(\widehat{c_1^-}) + \log \pi(\widehat{c_1^-}|y) + \log v(\widehat{c_2^-}) + \log \pi(\widehat{c_2^-}) \tag{25}$$

or equivalently

$$\log \pi(\widehat{c}^+|y) - \log \pi(\widehat{c_1^-}|y) - \log \pi(\widehat{c_2^-}|y) \geq \zeta \tag{26}$$

where

$$\zeta = \log v(\widehat{c_1^-}) + \log v(\widehat{c_2^-}) - \log v(\widehat{c}^+) \tag{27}$$

Whenever the parameter κ is large this property provides a vehicle for efficiently comparing the efficacy of adjacent $I-$simple partitions. Note that any optimal simple partition $\widehat{C}$ will maximize or equivalently

$$\mathbb{E}(\kappa^{-n_I}\{\overline{U}\left(\widehat{C}|C\right) + 1\}) = \overline{U}_I(\widehat{C}) + 0(\kappa^{-1}) \tag{28}$$

So if

$$\log \pi(\widehat{c}^+|y) - \log \pi(\widehat{c_1^-}|y) - \log \pi(\widehat{c_2^-}|y) > \zeta \tag{29}$$

then there is a κ for which $\overline{U}(\widehat{C}^+) > \overline{U}(\widehat{C}^-)$ and conversely if

$$\log \pi(\widehat{c}^+|y) - \log \pi(\widehat{c_1^-}|y) - \log \pi(\widehat{c_2^-}|y) < \zeta \tag{30}$$

then there is a κ for which $\overline{U}(\widehat{C}^+) < \overline{U}(\widehat{C}^-)$.

Note that under this subfamily of utilities we do not need to assume that C is $I-$simple, just that we only search over $\widehat{C}$ that are $I-$simple. This is because, if this $\widehat{C}$ is $I-$simple but C is not, then $C \neq \widehat{C}$ which - unlike in the more general scenario - in turn implies $U_I(\widehat{C}|C, \boldsymbol{\theta}(C)) = 0$.

4.2 Relationships between Product Utility and MAP

The implications of the results are the following:

1. From the comments in Section 3.2, to find the utility maximising partition we need only find the utility maximising partition over the potentially interesting genes I.

2. Under the product utility function, like the MAP score, the expected utility score decomposes making it possible to use simple standard search algorithms to explore the space for high scoring partitions.

From (26) if the combined cluster in the coarser partition has more interesting expected values than the two smaller clusters in the finer partition then we are more prepared to choose the coarser partition than under MAP. In this sense the local algorithm associated with a product utility can be seen as exactly a MAP search but

over the genes in I and with adjusted priors over the partition space: the original prior cohesions $\pi_0(\widehat{c}^+), \pi_0(\widehat{c_1^-}), \pi_0(\widehat{c_2^-})$ are simply replaced by the adjusted prior cohesions

$$\begin{aligned} \pi_0'(\widehat{c}^+|y) &= v(\widehat{c}^+)\pi_0(\widehat{c}^+|y) \\ \pi_0'(\widehat{c_1^-}|y) &= v(\widehat{c_1^-})\pi_0(\widehat{c_1^-}|y) \\ \pi_0'(\widehat{c_2^-}|y) &= v(\widehat{c_2^-})\pi_0(\widehat{c_2^-}|y) \end{aligned} \tag{31}$$

So, from an algorithmic perspective, searching for a simple partition maximising $\overline{U}_I(C)$ is almost identical to searching for a MAP model over the subclass of potentially interesting genes, except that the most interesting clusters are given a higher prior weight than the less interesting ones. To simplify our notation henceforth write c for $\widehat{c}$.

It is interesting to note that under appropriate conditions we can now find prior densities and $U_I \in \mathbb{V}_I$ where an optimal partition C under U_I is a MAP optimal partition on I. Thus assume $\phi(c)$ are balanced. The weights defined in (22) are then of the form

$$v(c) = n(c) \int w_c(\boldsymbol{\theta}_c)^{n(c)} \pi_c(\boldsymbol{\theta}_c|y)\, d\boldsymbol{\theta}_c \tag{32}$$

Second assume that the marginal utilities on the parameters are indicators so that when $\boldsymbol{\theta}_c \in \Psi$ where Ψ is a particular region of the parameter space the scientist is satisfied whilst otherwise she is not. Then

$$w_c(\boldsymbol{\theta}_c) = \begin{cases} 1 & \text{when } \boldsymbol{\theta}_c \in \Psi \\ 0 & \text{when } \boldsymbol{\theta}_c \notin \Psi \end{cases} \tag{33}$$

Under these conditions the threshold ζ defined in (27) can be written as

$$\zeta = \log P(\boldsymbol{\theta}_{c_1^-} \in \Psi|y) + \log P(\boldsymbol{\theta}_{c_2^-} \in \Psi|y) - \log P(\boldsymbol{\theta}_{c^+} \in \Psi|y) \tag{34}$$

In particular if we assume we have certainty ,i.e.

$$P(\boldsymbol{\theta}_{c_1^-} \in \Psi|y) = P(\boldsymbol{\theta}_{c_2^-} \in \Psi|y) = P(\boldsymbol{\theta}_{c^+} \in \Psi|y) = 1 \tag{35}$$

then $\zeta = 0$. So we recover MAP search but now restricted to I rather than Ω.

It is common practice in this context to first preselect genes that lie in a set I and then search for an optimal partition using MAP. So note that with the assumptions above this is a specific case of our method. Therefore our development above can be seen as providing a formal check about whether or not clustering combined with a particular preselection technique is valid and it also provides a way of adjusting this procedure when this is not so.

Note that we have shown that the preselection method is valid if the generating partition C is I-simple. When the scientist is not looking for specific structures, the sorts of routine preselection methods commonly used may well often be plausible. However when the scientist has strong beliefs about what shapes of data she is looking for, routine preselection will often not be consistent with the I-simple hypothesis. Furthermore, I

will certainly be larger than it is needed for the analysis and so the search unnecessarily computational hungry.

Summarising we therefore have the following result.

Theorem 4.2. *The relative score between adjacent $I-$simple partitions under a $U \in \mathbb{V}$ (19) score and cohesion priors is the relative MAP score over all partitions of I using the adjusted cohesion priors over the partition space given by (31). In the particular case when conditions (32), (33) and (35) hold then this relative product utility score over I-simple partitions is exactly the MAP score on all partitions of I using the original priors on the partition space.*

Thus in the very special case when (32), (33) and (35) hold the optimal MAP partition found by local search on I will also be optimal under product utility score over the space of simple partitions. It is simply that there are other optimal partitions under the product utility: namely those that differ from the MAP partition but only on the clustering of the uninteresting genes. The fact that there are so many more optimal partitions under product utility means that we are more likely, with an efficacious search algorithm, to find a high scoring partition more quickly. So when $I = \Omega$ our utility based search inherits all the search efficiency of local MAP search, whilst when $I \subset \Omega$ and we are content to search only for $I-$simple partitions, our search algorithm can focus on partitions optimal with respect to I. Then, in fact, the utility based search is *much quicker* than MAP.

We also can conclude that if $I \subseteq J$ and under U_J a $J-$simple locally optimal partition C_J^* is also $I-$simple then C_J^* is also locally optimal under the utility function U_I. In this sense, if we include too many genes in our potentially interesting set this will affect the efficiency of our search but not the optimality. So there is a robustness to mispecification of the set I, provided we err on the side of caution and include genes in I if we think they might be interesting. An illustration of this methodology is given in the second example of Section 5.3.

4.3 Robustness of the utility weighted score

Of course for most statistical models the certainty condition (35) will hold at best only approximately. Thus suppose that the scientist's utility has the form given by (32) and (33) but that we only know that, for all $c \in I$, $P(\boldsymbol{\theta}_c \in \Psi | y) \geq 1 - \alpha$ for some small value of α. If the interesting genes are discovered by thresholding then this rather than (35) may well be the type of condition we might have (see Section 5). Note that

$$-2\alpha \doteq 2\log(1-\alpha) \leq \zeta \leq -\log(1-\alpha) \doteq \alpha \tag{36}$$

It follows that when α is small and $I = \Omega$ then in this scenario our utility based search will closely approximate MAP search. So the method only performs significantly differently from MAP search in this context when either at least some of the clusters have weights α that are not small or when $I \subset \Omega$ which will be illustrated below.

Suppose therefore that, under the notation above

$$P(\boldsymbol{\theta}_{c_1^-} \in \Psi|y) \doteq P(\boldsymbol{\theta}_{c_2^-} \in \Psi|y) \doteq P(\boldsymbol{\theta}_{c^+} \in \Psi|y) = 1-\alpha \tag{37}$$

Then $\zeta \doteq \log(1-\alpha)$ would mean that for large α we would combine clusters much more often than under MAP: i.e. the partition will be coarser over the less interesting genes in I.

4.4 Some practical issues

With balanced criterion weights we have that

$$v(c) = n(c) \int u_c^0(\boldsymbol{\theta}_c)\pi_c\left(\boldsymbol{\theta}_c|y\right) d\boldsymbol{\theta}_c \tag{38}$$

where $u_c^0(\boldsymbol{\theta}_c) = w_c(\boldsymbol{\theta}_c)^{n(c)}$. Note that $u_c^0(\boldsymbol{\theta}_c) = w_c(\boldsymbol{\theta}_c)$ if $w_c(\boldsymbol{\theta}_c)$ is an indicator function. In order to implement our search algorithm to be comparably fast with MAP search we need to be able to quickly evaluate $v(c)$. This then provides the thresholds ζ determining whether or not we move to an adjacent partition. We could approximate this function using summary statistics calculated already and so obtain an approximately optimal partition. Alternatively we could try to find functions $u_c^0(\boldsymbol{\theta}_c)$ which on the one hand reflect the preferences of the biologist and on the other admit the explicit calculation of $v(c)$.

In our running example we need to find expedient $u_c^0(\boldsymbol{\theta}_c)$ - when $\pi(\boldsymbol{\theta}_c|y)$ which has a product Gaussian - Inverse Gamma form - making $v(c)$ an explicit function of the hyperparameters of $\pi(\boldsymbol{\theta}_c|y)$. The second alternative is clearly more elegant, but we have found that the first option is more flexible and appears to be robust to the approximations we take.

5 Examples

To investigate the efficacy of this method we studied the circadian rhythms of the plant *Arabidopsis thaliana*. The experimental results were provided by co-authors Kieron D. Edwards and Andrew J. Millar and have been published in Edwards et al. (2006), although the analysis performed below is more refined than the original. We compare standard MAP methods used with AHC local search to our expected utility based search using adapted AHC on the same datasets.

We will illustrate our proposed method and its efficacy on a smaller example of 18 observations. Then we will show how an approximation of the methodology presented can be used on a larger example of tens of thousands of observations. For simplicity we will concentrate on an analysis where conditions (32), (33) and (35) are assumed to hold and we use an approximation rather than an exact evaluation of (38).

Table 1: The score of the best partition of the interesting genes obtained with direct AHC and AHC but applied to potentially interesting gene only.

	$\Sigma(C)$
Direct AHC	64.896
AHC on I	68.295

5.1 Data

The gene expression of 22,810 genes was measured by Affymetrix microarrays at 13 time points over two days. The aim was to identify the genes (of order 1,000) which may be connected with the circadian clock of the plant. Constant white light was shone on the plants for 26 hours before the first microarray was taken, with samples every four hours. The light remained on for the rest of the time course. Thus, there are two cycles of data (13 time points) for each of the 22,810 genes available on the *Arabidopsis* microarray chip. Subjective dawn occurs at about the 24th and 48th hours – this was when the plant has been trained to expect light after 12 hours of darkness. An exposition of the whole dataset, together with a discussion of its biological significance is given in Edwards et al. (2006) and subsequently by Michael et al. (2008).

The cluster profiles at time t, $y(t)$, over a 48 hour time course are given by

$$y(t) = \beta_0 + \sum_{i=1}^{6} [\beta_{2i-1} \cos(2\pi it/48) + \beta_{2i} \sin(2\pi it/48)] \tag{39}$$

5.2 A simple example of how direct use of AHC fails

Here 15 circadian genes from the dataset above, were contaminated with 3 outliers. Using AHC on the 18 genes produces two clusters (see Figure 2). The second cluster contains all potentially interesting genes.

To compare this with our utility based approach we simply specified our set of interesting genes I as those whose individual first harmonics are a larger than the expected proportion of the total variation, here set to 0.25. Preselecting the set I in this coarse way identifies the 15 genes in the second cluster in Figure 1. However now using AHC on I leads to the further discrimination of the 15 genes into the two clusters in Figure 2.

It is easy to see that both in terms of their Bayes factor scores and visually these new clusters discriminate profiles much better than AHC used directly. AHC is disrupted by outliers in larger problems in similar ways. See Smith et al. (2008) for reasons for this. When potentially interesting profiles can be defined then even using a crude filter like the one illustrated above and then using a simple local search algorithm like AHC on I can greatly enhance the discovery process and classify interesting genes more precisely. We have seen earlier that proceeding in this way is formally justified provided the scientist has a utility as defined in Section 4 with equal utility weights.

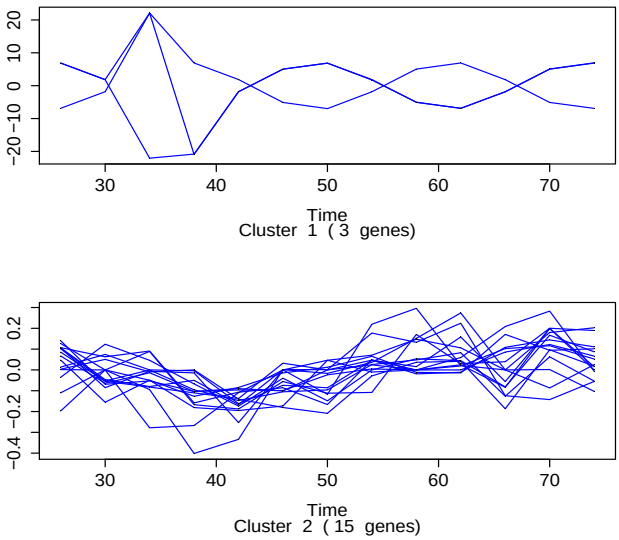

Figure 1: Clusters obtained on 18 genes of *Arabidopsis thaliana* running AHC only once on the whole dataset ($\Sigma(C) = 64.896$). The y-axis is the log of gene expression. Note the different y-axis scale for the two clusters. This is the partition with the lowest score.

5.3 A simple approximate guided learning algorithm

We have shown in the previous example that AHC does not always succeed in identifying the best scoring partition and that our proposed utility method can enhance the clustering algorithm by formally allowing for the search to focus on interesting areas of the partition space. This is particularly important when dealing with high dimensional datasets, such as the whole dataset of 22,810 gene expression profiles of Arabidopsis.

It is usual to prefilter using either a simple expression threshold (Zhou et al. 2006) or a naive simple filter, such as the first harmonic in the Cosopt software (Straume 2004) or prefiltering techniques as in Eisen et al. (1998), Tamayo et al. (1999), Wakefield et al. (2003) and Zhou et al. (2006). However in our context we found that prefiltering in this way removed a high proportion of genes whose profiles looked interesting to the biologist, because it was special shapes of harmonic, often non sinusoidal, profiles and sometimes relatively lowly expressed profiles which experience had suggested had biological regulatory importance. By performing the more refined filters of preclustering we were able to reduce the variance of estimated flexibly shaped profiles when these were replicated, helping to ensure that circadian but lowly expressed genes appeared in the set I we subsequentially searched.

First of all, we preclustered using the Bayes factors associated to a full Bayesian clustering algorithm on subsets of genes using the usual conjugate analyses by Heard et al. (2006) but adapted to a Fourier domain. We then treated the cluster parameter distributions as providing rough estimates of the profile of the *individual* genes contained

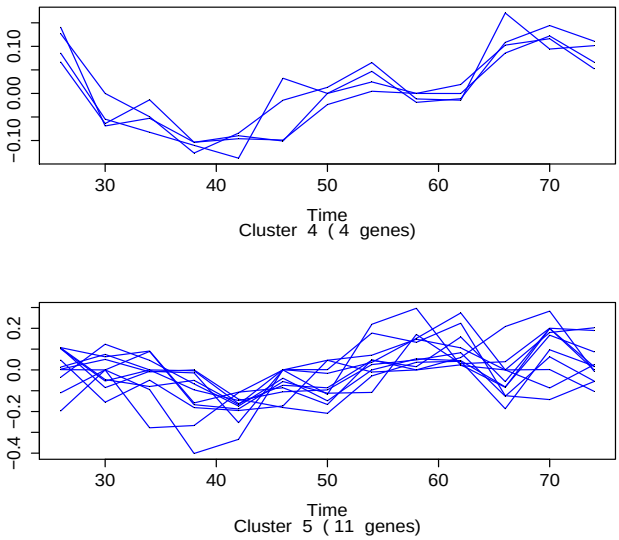

Figure 2: Running AHC again on the genes in I we do not search the partition space around the 3 outliers, but we find a higher scoring partition for the other 15 genes. The score of this new partition is $\Sigma(C_{\text{Iterative}}) = 68.295$.

in each particular cluster. We noticed that although cluster containment could be very sensitive to the setting of hyperparameters the estimates of individual gene profiles was remarkably robust to our hyperparameter settings: see Appendix 6. The only exception to this was that, because of certain technical difficulties described in Smith et al. (2008), these algorithms occasionally produced 'junk' clusters containing many genes with highly or moderately highly expressed, but heterogenous, profiles. The cautious approach advocated in Section 4.2 therefore suggested we included these genes into the class I of interesting genes. So the set I constituted genes with well estimated profile means in interesting areas of the parameter space together with genes whose profile estimated variance was large.

There were several options for defining regions of potential interest. One was to use the posterior distribution of a measure of the interestingness of a cluster profile being greater than a given threshold using the MAP estimate of each gene. In our particular context biological expert judgement suggested that an interesting cluster is one whose second harmonic is high relative to the third, fourth, fifth and sixth harmonics.We define the *second harmonic ratio* (SHR) as

$$\text{SHR} = \left(\beta_3^2 + \beta_4^2\right)^{\frac{1}{2}} \Bigg/ \sum_{i=1,3,5,7,9,11} \left(\beta_i^2 + \beta_{i+1}^2\right)^{\frac{1}{2}} \tag{40}$$

Because the distribution of this measure was not in closed form, provided the estimated variance of the regression parameters was not large, we approximated this by substituting the posterior means for their actual value in the thresholding formulae, as suggested

in Section 4.4.

Once AHC has been used to discriminate the set I it is possible to use more refined search techniques on smaller sets. However, for the purpose of this illustration in this paper we simply ran AHC again but now restricted to I. By doing this, we found that the contribution to the marginal likelihood over the set I of the final pass was much greater than that associated with the marginal likelihood over interesting genes found in the final run because outlying genes were largely sieved out through the iteration process. From Section 4 this means that the utility score for these new partitions was also greater. Full results of this final pass are given in Appendix 6, where the clustering can be seen to be tight.

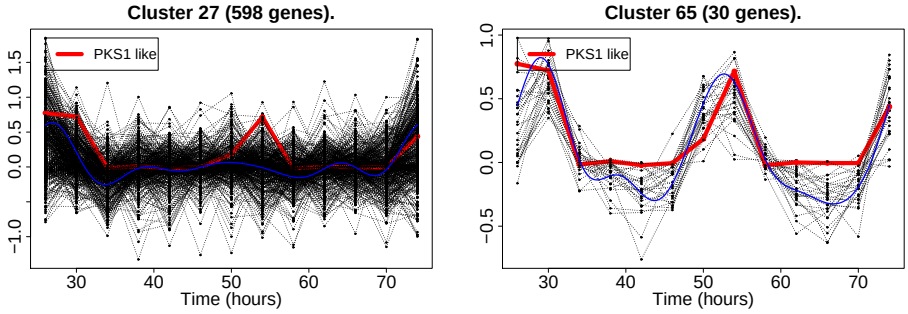

Figure 3: Demonstration of the advantages of the utility-based search algorithm. On the left is a potentially not interesting cluster from the penultimate step containing PKS1-like which biologists believe may be involved in the clock. The cluster on the right shows that PKS1-like ends up reclassified as a potentially circadian cluster after AHC has been reused on the subset I alone.

An example of the practical as well as theoretical usefulness of our utility based algorithm is illustrated in figure 3 using the unguided standard AHC method, whatever the values of the hyperparameters, the regulatory PKS1-like gene was always classified in a high variance no signal cluster like the one depicted in the first graph of figure 3. However, by first identifying the subset I of interesting genes the profile of this possibly regulatory gene is reclassified into a new cluster which is clearly circadian. Potentially useful possible homologues of the PKS1-like gene can now be identified as those genes whose profile lie in this cluster.

6 Discussion

Guided Bayesian clustering methods like the simple one described here clearly enhance the performance of Bayesian clustering algorithms for longitudinal time series. Our proposed methods can explore much larger relevant regions of the partition space and provide a useful, practical and formally defensible tool for the search of high dimensional partition spaces where the units in the partition are not exchangeable. Note that our techniques apply outside the narrow context of clustering gene profiles. Any clustering

of large numbers of units can benefit from the approach discussed above, provided that the domain expert can be specific enough about her priorities to specify relative utility weights.

Of course there are significant further improvements that can be made to the methodology above. We are currently exploring such avenues. First we are examining the effects of using more refined local search techniques on the final stage of the process illustrated in the second example above, allowing splits and long jumps, currently used in other domains (Chipman et al. 2002). A second improvement is to weight the interesting genes, as described in the last section, rather than use a simple indicator discriminant. It is easy to do this if we approximate using the continuous score on the SHR obtained as a function of the means of the parameters in the penultimate iteration. Moreover, in the example we used an approximated guided algorithm, but those approximations are unnecessary for some expedient choices of utility functions. Instead of SHR, measures with a known distribution could have been used for precision rather than speed. Our results so far, though, showed that the practical gain in such exact methods, although measurable, was not great. Finally, we are investigating the effects of using heterogeneous versions of other partition priors discussed in Booth et al. (2008), Crowley (1997) and McCullagh and Yang (2006). We will report our findings in a later paper.

Appendix

1. Plots and classification of the clustering

In this appendix we give the results of our I-MAP optimally found clusters of the final AHC partition search. Because of the speed of our method, many iterations could be performed over a grid of hyperparameters. The results given below correspond to the clustering giving the highest score with the choice of variance matrix $V = vI$ where $v = 0.498$.

After the clustering process, it was found helpful to the biologists to classify the posterior mean profiles into various shapes. The first five classifications cover the clusters identified as circadian over both 24-hour periods; the last five those that aren't. Types I, VI and VII are delineated by objective criteria whilst the remaining types are classified by eye. This is nevertheless useful as a guideline to the broad classes of behaviour that are displayed.

Further details on the analysis can be found in Anderson et al. (2006).

I. Sinusoidal: those clusters with SHR>0.65 and more than 11 genes.

II. Sharply rising then sharply falling.

III. Sharply falling then sharply rising.

IV. Sharply rising then drifting back to zero.

V. Sharply falling then drifting back to zero.

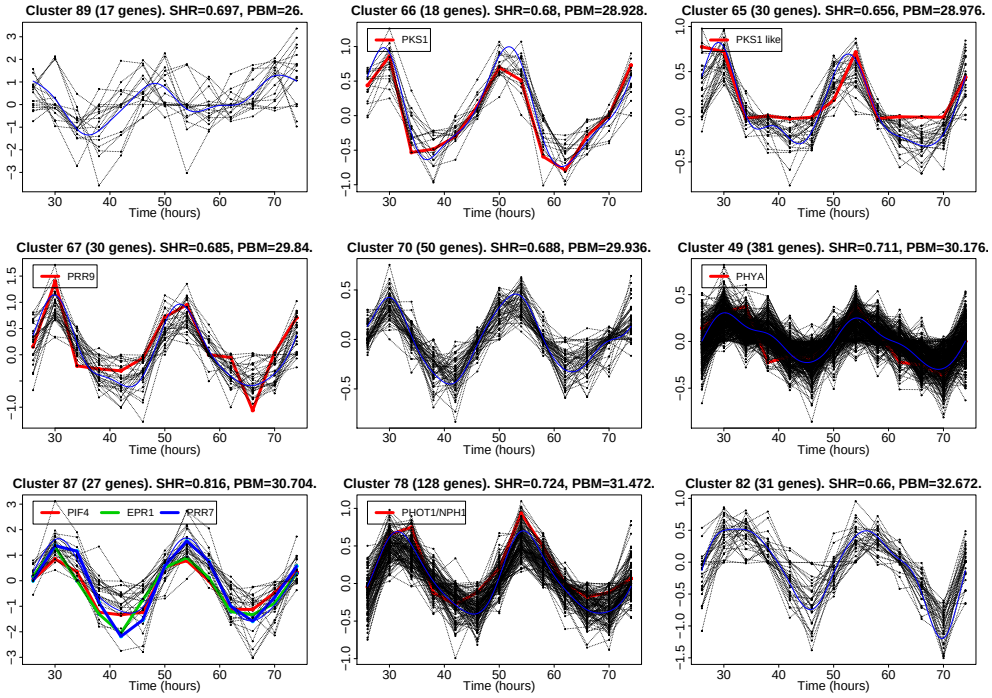

Figure 4: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

VI. Clusters classified as potentially not interesting by the algorithm.

VII. Clusters classified as O by the algorithm.

VIII. Potentially circadian, but not accurately repeated: clusters with a peak or trough in one 24 hour period, but not in the other.

IX. Outliers: clusters containing less than 11 genes.

X. 'Junk': clusters with expressions close to zero and non-circadian profiles.

The profiles of each of the 100 clusters identified among the interesting genes are shown in figures 4 to 19 classified in order according to the ten types above. Within each type, the clusters are sorted by their phase by maximum (the maximum value of the posterior mean in the first 24 hours). The second harmonic ratio (SHR) and phase by maximum (PBM) are given on each plot.

Figures 20 and 21 illustrate the robustness of the initial gene estimates to mispecification of hyperparameters. The clustering is very different when $v = 10,000$ rather

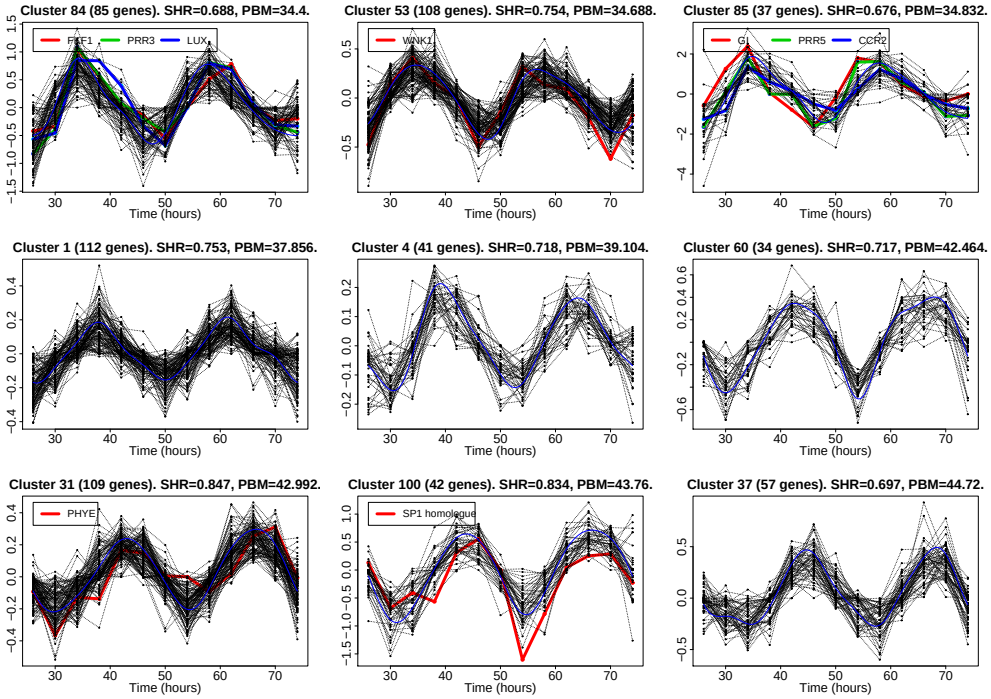

Figure 5: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

than $v = 0.498$ giving us far fewer, but larger, clusters (34 instead of 100) so that we get intrinsically different solutions. Despite this, the estimated profiles of most genes do not radically differ under changes in v. This means that the sets of genes identified to have interesting profiles do not change greatly over large ranges of v. Furthermore, genes having similar profiles remain close.

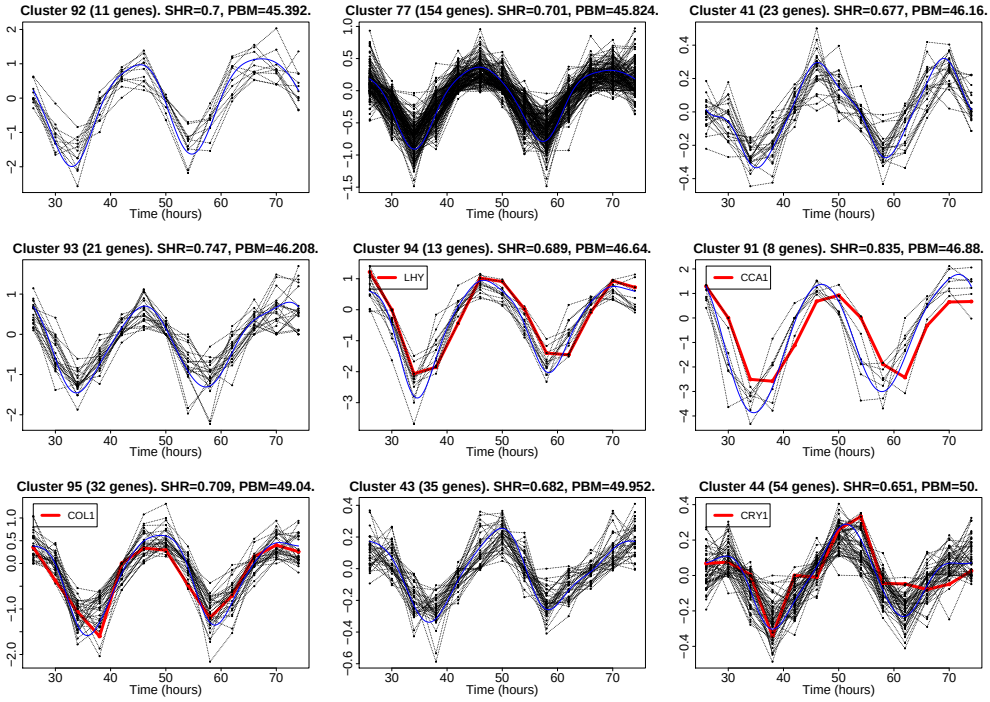

Figure 6: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

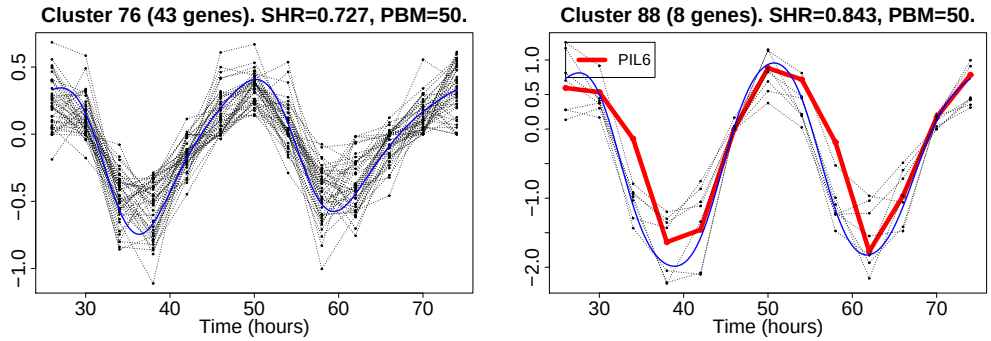

Figure 7: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

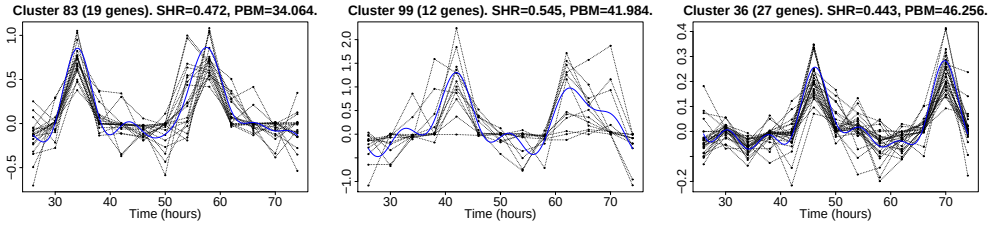

Figure 8: Type II clusters: sharply rising then sharply falling. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

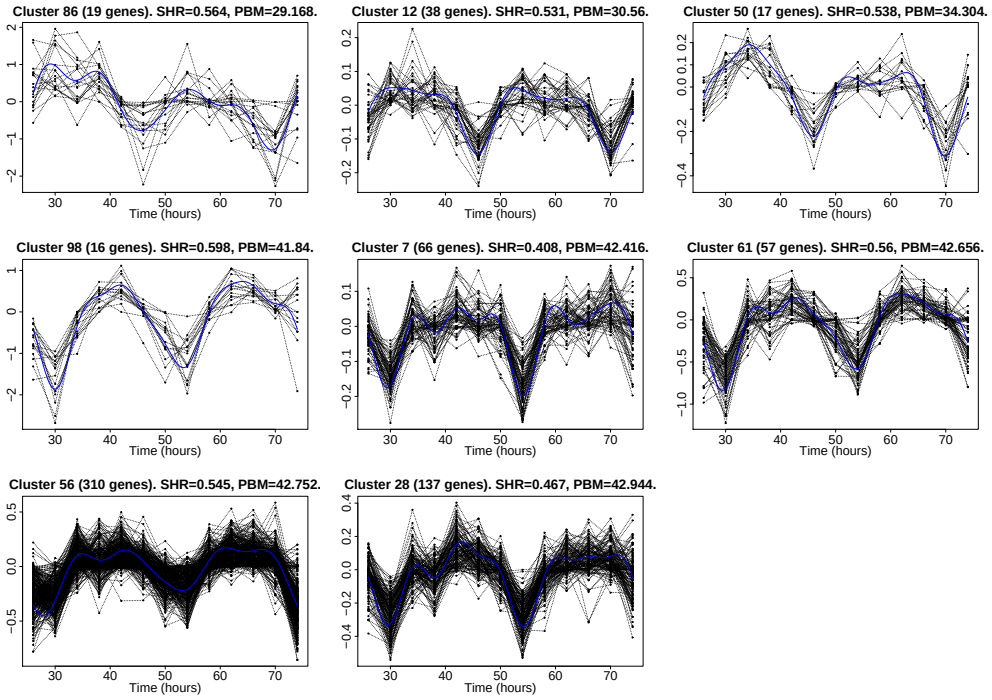

Figure 9: Type III clusters: sharply falling then sharply rising. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

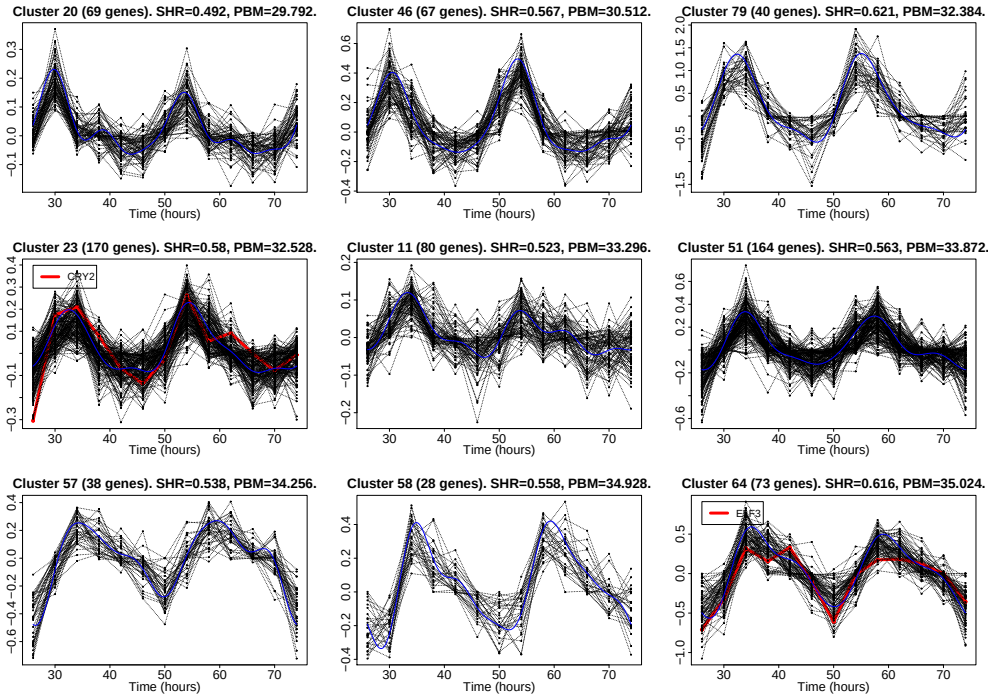

Figure 10: Type IV clusters: sharply rising then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

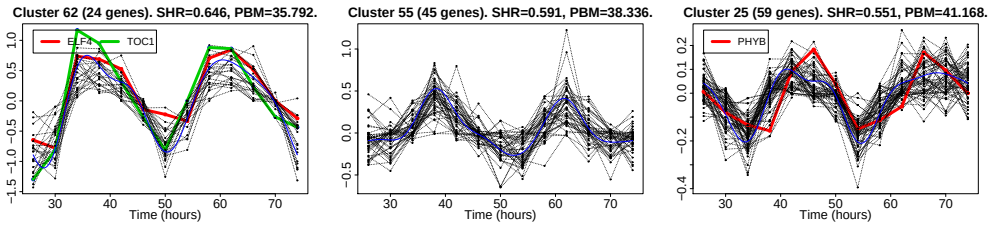

Figure 11: Type IV clusters: sharply rising then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

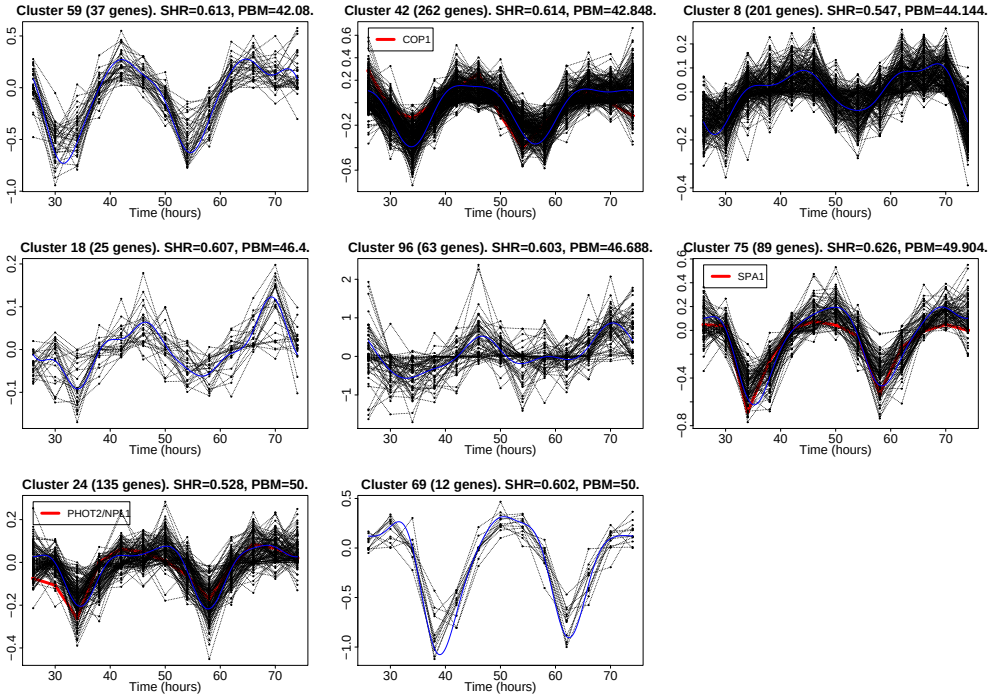

Figure 12: Type V clusters: sharply falling then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

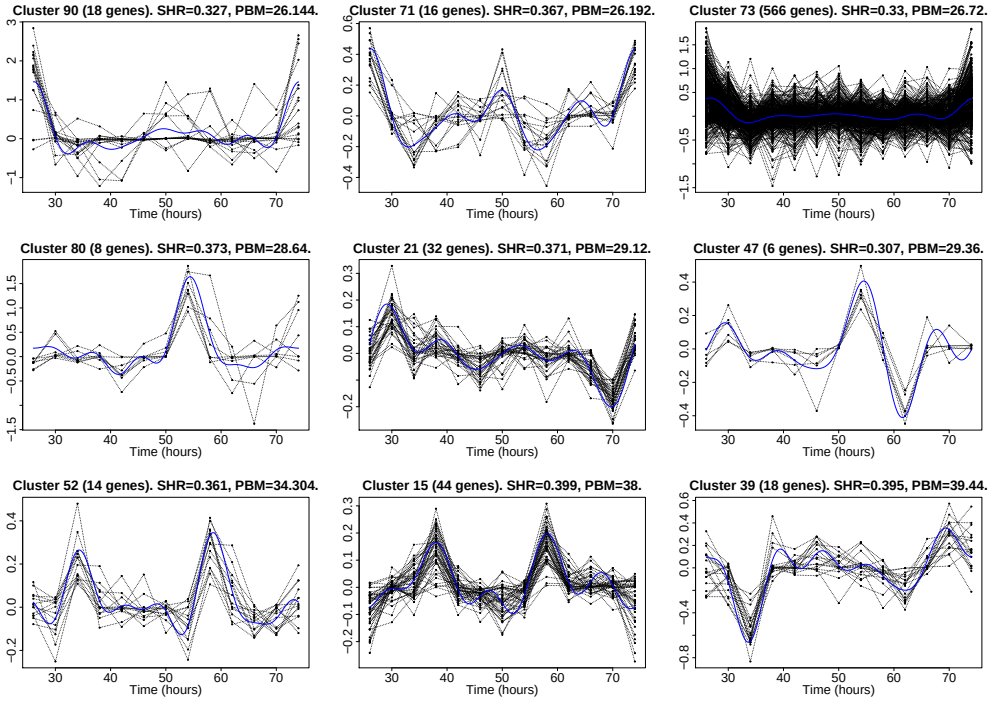

Figure 13: Type VI clusters: potentially not interesting. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

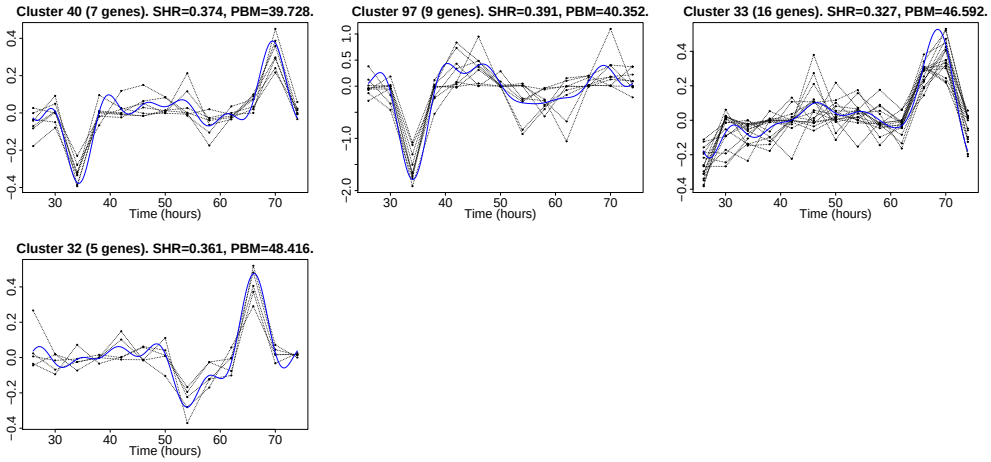

Figure 14: Type VI clusters: potentially not interesting. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

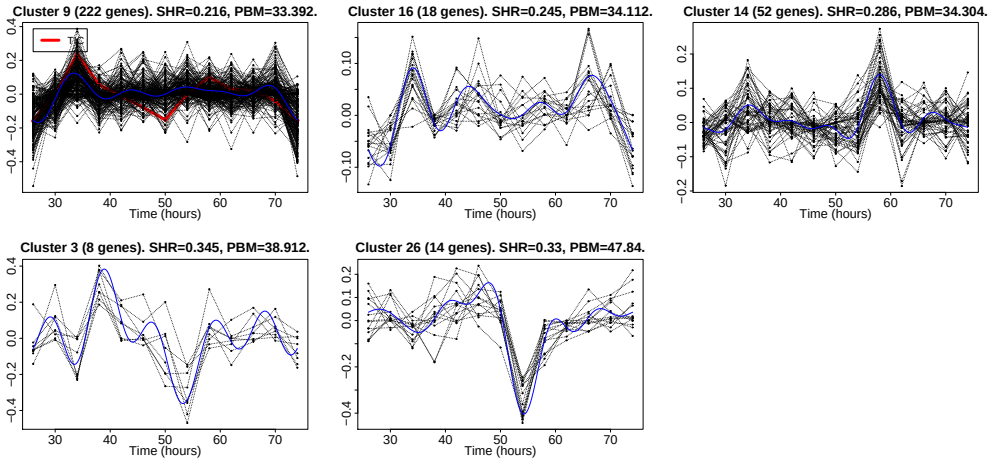

Figure 15: Type VII clusters: other. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

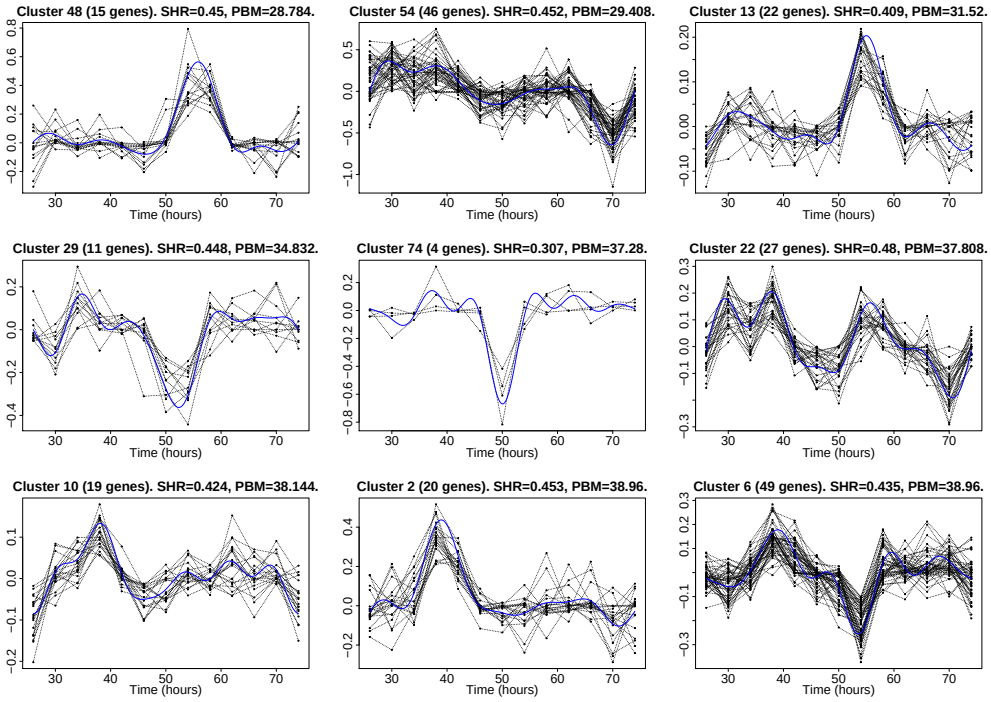

Figure 16: Type VIII clusters: potentially circadian, but not repeated. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

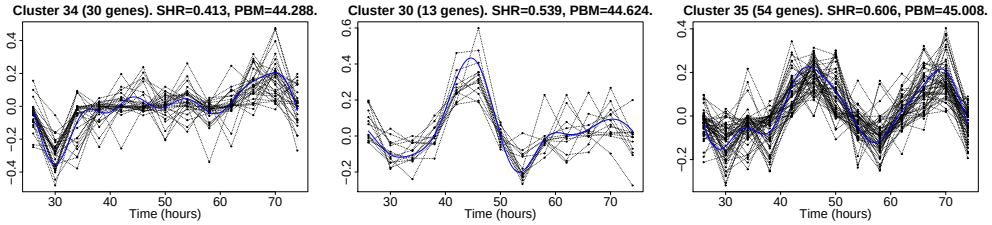

Figure 17: Type VIII clusters: potentially circadian, but not repeated. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

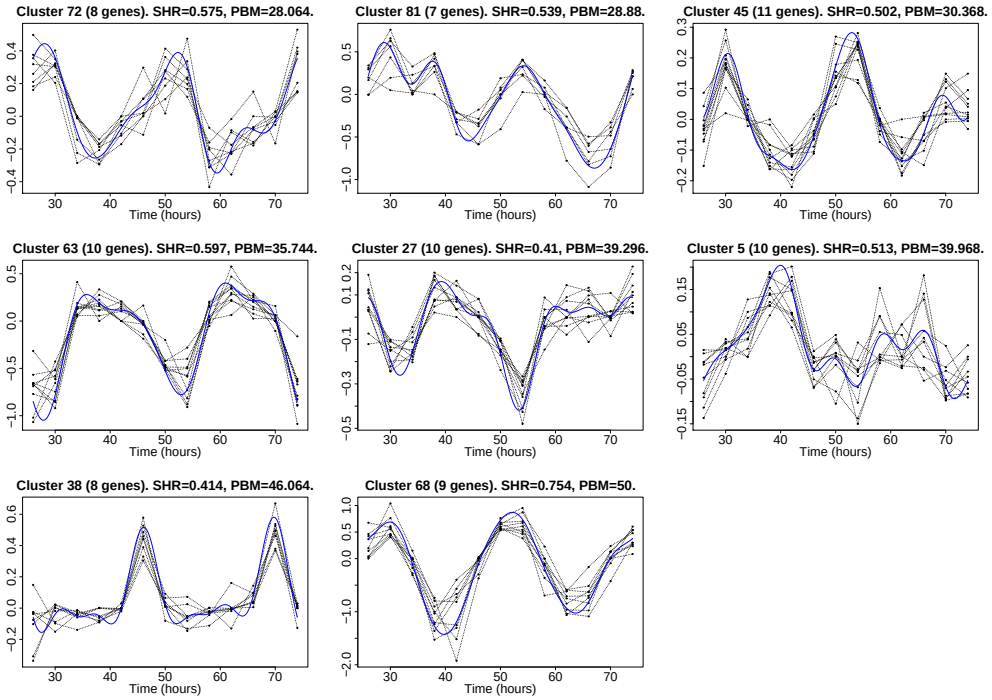

Figure 18: Type IX clusters: outliers. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

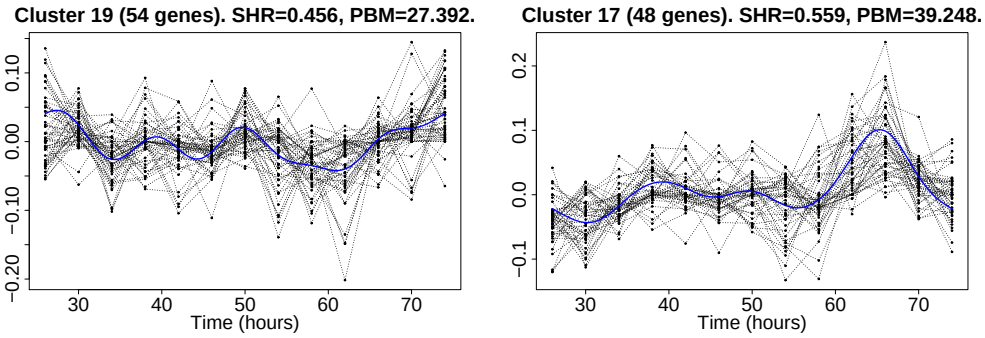

Figure 19: Type X clusters: not interesting. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

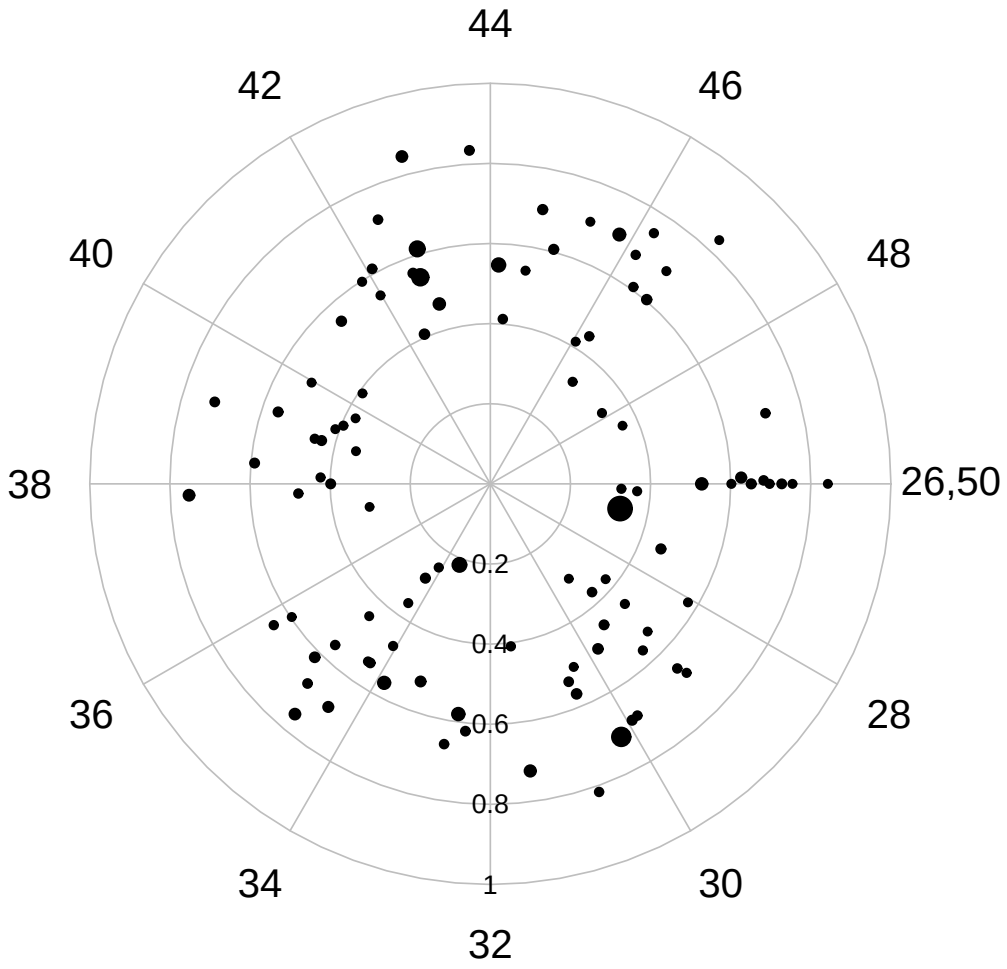

Figure 20: Phase plot of the clusters of the final gene set with $v = 0.498$. Each dot is one cluster, its radius is proportional to the number of genes it contains. Its distance from the origin gives its second harmonic ratio, and the angle indicates the phase by maximum of the posterior mean profile.

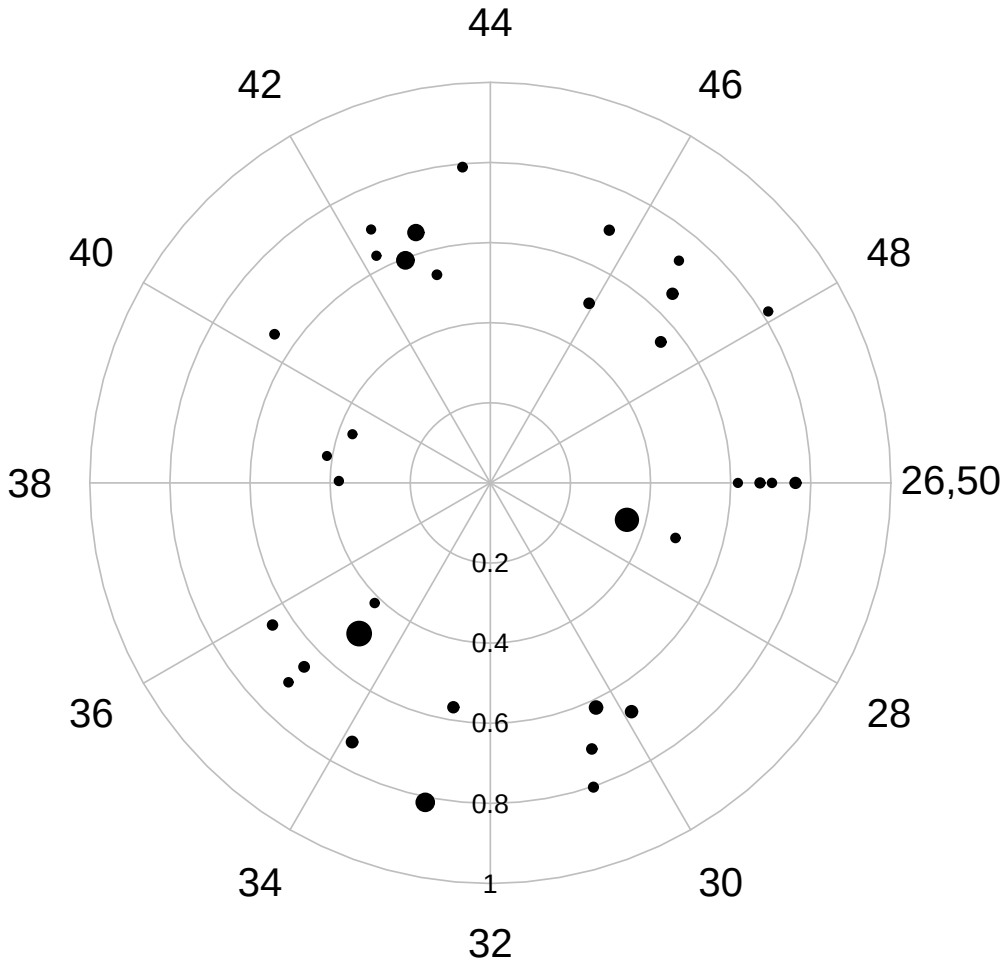

Figure 21: Phase plot of the clusters of the final gene set with $v = 10,000$. Each dot is one cluster, its radius is proportional to the number of genes it contains. Its distance from the origin gives its second harmonic ratio, and the angle indicates the phase by maximum of the posterior mean profile. The structure is broadly similar to that of figure 20.

References

Anderson, P. E., Smith, J. Q., Edwards, K. D., and Millar, A. J. (2006). "Guided Conjugate Bayesian Clustering for Uncovering Rhytmically expressed Genes." *CRISM Working Paper*, (07). 556

Banfield, J. D. and Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics*, 49(3): 803–821. 540

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). "Clustering Gene Expression Patterns." *Journal of Computational Biology*, 6(3–4): 281–297. 540

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. 540

Booth, J. G., Casella, G., and Hobert, J. P. (2008). "Clustering using objective functions and stochastic search." *Journal of the Royal Statistical Society, Series B*, 70(1): 119–139. 556

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). "Bayesian treed models." *Machine Learning*, 48(1–3): 299–320. 556

Crowley, E. M. (1997). "Product Partition Models for Normal Means." *Journal of the American Statistical Association*, 92(437): 192–198. 556

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. John Wiley and Sons. 540, 541

Edwards, K. D., Anderson, P. E., Hall, A., Salathia, N. S., Locke, J. C. W., Lynn, J. R., Straume, M., Smith, J. Q., and Millar, A. J. (2006). "FLOWERING LOCUS C Mediates Natural Variation in the High-Temperature Response of the *Arabidopsis* Circadian Clock." *The Plant Cell*, 18: 639–650. 541, 551, 552

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." *Proceedings of the National Academy of Sciences*, 95(25): 14863–14868. 553

Fraley, C. and Raftery, A. E. (1998). "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis." *The Computer Journal*, 41: 578–588. 539, 540

French, S. and Rios Insua, D. (2000). *Statistical Decision Theory*. London: Arnold. 543

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves." *Journal of the American Statistical Association*, 101(473): 18–29. 539, 540, 541, 553

Keeney, R. and Raiffa, H. (1976). *Decision with multiple objectives: Preferences and value tradeoffs*. New York: John Wiley & Sons. 543

Keeney, R. and von Winterfeldt, D. (2007). "Practical Value Models." In Edwards, W., Miles, R. F., and von Winterfeldt, D. (eds.), *Advances in Decision Analysis: From Foundations to Applications*, 232–252. Cambridge University Press. 543

Luan, Y. and Li, H. (2003). "Clustering of time-course gene expression data using a mixed-effects model with B-splines." *Bioinformatics*, 19(4): 474–482. 540

McCullagh, P. and Yang, J. (2006). "Stochastic classification models." In *Proceedings International Congress of Mathematicians*, volume III, 669–686. 556

Michael, T., Mockler, T., Breton, G., McEntee, C., Byer, A., Trout, J., Hazen, S., Shen, R., Priest, H., Sullivan, C., Givan, S., Yanovsky, M., Hong, F., Kay, S., and Chory, J. (2008). "Network Discovery Pipeline Elucidates Conserved Time-of-Day–Specific cis-Regulatory Modules." *PLoS Genetics*, 4(2): e14. 552

O'Hagan, A. and Forster, J. (2004). *Bayesian Inference: Kendall's Advanced Theory of Statistics*. Arnold, second edition. 541

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). "Cluster Analysis of Gene Expression Dynamics." *Proceedings of the National Academy of Sciences of the United States of America*, 99(14): 9121–9126. 540

Ray, S. and Mallick, B. (2006). "Functional clustering by Bayesian wavelet methods." *J. Royal Statist. Soc.: Series B*, 68(2): 305–332. 541

Smith, J. Q., Anderson, P. E., and Liverani, S. (2008). "Separation Measures and the Geometry of Bayes Factor Selection for Classification." *Journal of the Royal Statistical Society, Series B*, 70(5): 957–980. 540, 541, 542, 552, 554

Straume, M. (2004). "DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning." *Methods Enzymol*, 383: 149–66. 553

Tai, Y. and Speed, T. (2006). "A multivariate empirical Bayes statistic for replicated microarray time course data." *Annals of Statistics*, 34(5): 2387–2412.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." *Proceedings of the National Academy of Sciences*, 96(6): 2907–2912. 553

Tatman, J. and Shachter, R. (1990). "Dynamic programming and influence diagrams." *Systems, Man and Cybernetics, IEEE Transactions on*, 20(2): 365–379. 540

Wakefield, J., Zhou, C., and Self, S. (2003). "Modelling gene expression over time: curve clustering with informative prior distributions." *Bayesian Statistics*, 7: 721–732. 540, 553

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). "Model-based clustering and data transformations for gene expression data." *Bioinformatics*, 17(10): 977–987. 540

Zhou, C., Wakefield, J. C., and Breeden, L. L. (2006). "Bayesian Analysis of Cell-Cycle Gene Expression Data." In Do, K.-A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 177–200. Cambridge University Press. 540, 553

Acknowledgments

We thank Nick Heard and Chris Holmes for useful discussions. KDE's experimental work was supported by BBSRC grant G19886 to AJM. JQS and AJM are supported by the joint ANR/BBSRC Systems Biology initiative, award BBF0054661. The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and EPSRC, reference BB/D019621/1. SL and JQS thank EPSRC through CRiSM for financial support.