Clustering with Proportional Scaling

Jim Q. Smith†

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

Paul E. Anderson

Systems Biology Centre and Department of Statistics, University of Warwick, UK

Silvia Liverani

Department of Statistics, University of Warwick, UK

Summary. Conjugacy assumptions are often used in Bayesian selection over a partition because they allow the otherwise unfeasibly large model space to be searched very quickly. The implications of such models can be analysed algebraically. In this paper we use the explicit forms of the associated Bayes factors to demonstrate that such methods can be unstable under common settings of the associated hyperparameters. We then prove that the regions of instability can be removed by setting the hyperparameters in an unconventional way. Under this family of assignments we prove that model selection is determined by an implicit separation measure: a function of the hyperparameters and the sufficient statistics of clusters in a given partition. We show that this family of separation measures has desirable properties. The proposed methodology is illustrated through the selection of clusters of longitudinal gene expression profiles.

Keywords: Bayes factors, classification, non-conjugate analyses, g-priors, clustering, separation measures.

1. Introduction

When a model space is vast, it is often expedient to select a Bayesian model using conjugate priors; see for example Barry and Hartigan (1992) and Heard et al. (2006). The Bayes factors then have a simple algebraic form so the comparison of two models is then almost instantaneous. This makes search algorithms for models with high posterior probability in this huge partition space orders of magnitude faster than their numerical non-conjugate analogues.

In this paper we demonstrate that the explicit nature of this type of selection algorithm has another advantage. The properties and characteristics of the algorithm can be studied algebraically. In our particular case, its underlying geometry is linked with the well-studied behaviour of products of t-distributions; see for example O'Hagan and Le (1994), Chipman et al. (2001) and references therein. This enables us to explain not only how and why conjugate Bayesian model selection can break down under default settings of hyperparameters, but also to show that most of these apparent anomalies are removed if the hyperparameters are calibrated to plausible pre-posterior predictions, within a particular subfamily of these conjugate models.

†Author to whom correspondence should be addressed: j.q.smith@warwick.ac.uk

In the next section we briefly review the geometry of the types of products of t-densities which form the marginal likelihoods of this class. In section 3 we demonstrate how this geometry impinges on model selection based over partitions with particular emphasis on the methodology proposed in Heard et al. (2006). We illustrate how and why standard settings of hyperparameters can produce poor selection characteristics in section 4. In section 5 we derive explicit characterisations ensuring that Bayes factor selection prefers partitions that combine clusters when they are close with respect to a certain separation measure. In section 6 we illustrate these new settings in certain idealised contexts and in section 7 we examine properties of this implicit separation measure. This enable us to make a direct link between Bayes Factor selection and more conventional separation based clustering methods; see Chipman and Tibshirani (2006), Gordon (1999) and Hastie et al. (2001). We demonstrate that a partition, C_1, is preferred to another, C_2, (which is identical to C_1 except that two particular clusters in C_1 are combined into one cluster in C_2) if and only if the sufficient statistics of the two clusters in C_1 are different enough from one another in a certain, very natural, sense.

In a careful study of model selection over large spaces of linear models, Chipman et al. (2001) argue that hyperparameters should be set to make prior assumptions minimally influential. However, when selecting across a space of partition models, we argue that such a strategy is futile and all settings of hyperparameters have a different and strong effect on model selection over this domain. The analysis below allows us to adopt a proper Bayesian approach analogous to Garthwaite and Dickey (1992) which is straightforward in this domain. We demonstrate that it is simple to elicit values of hyperparameters within the class of proportional models so that they calibrate to pre-posterior predictions associated with the model space in a given context. It is also possible to demonstrate both analytically and numerically that these settings are robust to moderate misspecification. We begin the paper with some technical background.

2. A Simple Likelihood Ratio

2.1. Conjugate Bayesian estimation of profiles

Consider the Gaussian conjugate Bayesian regression model where $\boldsymbol{D} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ and $\boldsymbol{Y} = \text{vec}(\boldsymbol{D})$ satisfy

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

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)' \in \mathbb{R}^p$ and $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$ is a vector of independent error terms with $\sigma^2 > 0$. Note that $\boldsymbol{Y}_i \in \mathbb{R}^r$ 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

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

with

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

where $a, b > 0$ and V is a positive definite matrix. Throughout this paper we assume that $X = \mathbf{1}_n \otimes B$, where B is a known matrix, and that $X'X = nB'B$ is full rank. The Bayes factor associated with this model can then be calculated from its marginal likelihood $L(\boldsymbol{y})$,

see for example page 240 of Denison et al. (2002) and pages 308–12 of O'Hagan and Forster (2004). Thus

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

which can be written as

$$2\log L(\boldsymbol{y}) = 2l(\boldsymbol{y}) = K(V, a, b, n) - 2a^* \log(b + \gamma/2)$$

where

$$K(V, a, b, n) = 2\log\left(\left(\frac{1}{\pi}\right)^{nr/2} b^a \frac{|V^*|^{1/2}}{|V|^{1/2}} \frac{\Gamma(a^*)}{\Gamma(a)}\right).$$

Because $X'X$ is full rank the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of the mean vector $\boldsymbol{\beta}$ is uniquely defined and given by

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X\boldsymbol{y} = n^{-1}(B'B)^{-1}B'\boldsymbol{D}\mathbf{1}$$

and

$$\gamma = rn\widehat{\sigma}^2 + \widehat{\boldsymbol{\beta}}'(V + (X'X)^{-1})^{-1}\widehat{\boldsymbol{\beta}}$$

where $\widehat{\sigma}^2 = (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})/rn$ is the maximum likelihood estimate of σ^2.

Note that, as pointed out by Heard et al. (2006) p. 19, this hierarchical structure has the useful property of modeling the time dependance of the vector of observations in each profile $\boldsymbol{Y}$. For example, the Fourier basis B we later use in a running example allows us to model any individual profile so that its predictive distribution is an arbitrary weakly stationary process (see West et al., 1997).

2.2. *Comparing two regression profiles*

Define the observation vector $\boldsymbol{Y} = (\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})'$, where $\boldsymbol{Y}^{(j)} = (\boldsymbol{Y}_1^{(j)}, \ldots, \boldsymbol{Y}_{n_j}^{(j)})'$. The components $\{\boldsymbol{y}_i^{(j)} : 1 \le i \le n_j, j = 1, 2\}$ are profiles of a fixed length $r \ge p$ with

$$\boldsymbol{Y}_i^{(j)} = B\boldsymbol{\beta}_j + \varepsilon_i^{(j)}$$

where $\varepsilon_i^{(j)} \sim N(\mathbf{0}, \sigma_j^2 I)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\coprod_{i,j} \varepsilon_i^{(j)} | \boldsymbol{\beta}$ with $\coprod$ representing independence between random variables. Thus the profile vectors containing the longitudinal data on each unit, $\boldsymbol{Y}_i^{(j)}$, each follow the same linear model with a design matrix B of rank p. $\boldsymbol{Y}^{(j)}$ is a vector of length rn_j, $j = 1, 2$.

Let model M_s assume that the vectors $\boldsymbol{Y}^{(1)}$, $\boldsymbol{Y}^{(2)}$ are independent with $(\boldsymbol{\beta}_1, \sigma_1^2) \coprod (\boldsymbol{\beta}_2, \sigma_2^2)$, where $(\boldsymbol{\beta}_j, \sigma_j^2)$ is assumed to have the prior density $NIG(\mathbf{0}, V_j, a_j, b_j)$. Then, with the obvious extension of the notation given above, its log marginal likelihood $l_s(\boldsymbol{y})$ is given by

$$2l_s(\boldsymbol{y}) = \sum_{j=1,2} K(V_j, a_j, b_j, n_j) - 2\sum_{j=1,2} a_j^* \log(b_j + \gamma_s^{(j)}/2)$$

where

$$\gamma_s^{(j)} = rn_j\widehat{\sigma}_j^2 + \widehat{\boldsymbol{\beta}}_j'(V_j + n_j^{-1}(B'B)^{-1})^{-1}\widehat{\boldsymbol{\beta}}_j$$

and

$$\widehat{\boldsymbol{\beta}}_j = n_j^{-1}(B'B)^{-1}B'\boldsymbol{D}_j\mathbf{1}.$$

Now, compare the model M_s with a model M_t that assumes the vectors $\boldsymbol{Y}^{(1)}$, $\boldsymbol{Y}^{(2)}$ share the same parameter values. Under M_t, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $\sigma_1^2 = \sigma_2^2$ where $(\boldsymbol{\beta}_1, \sigma_1^2)$ has the prior density $NIG(\mathbf{0}, \overline{V}, \overline{a}, \overline{b})$. So the log marginal likelihood $l_t(\boldsymbol{y})$ of this model satisfies

$$2l_t(\boldsymbol{y}) = K(\overline{V}, \overline{a}, \overline{b}, n_{12}) - 2\overline{a}^* \log(\overline{b} + \gamma_t/2)$$

where $n_{12} = n_1 + n_2$ and

$$\gamma_t = rn_{12}\widehat{\sigma}^2 + \widehat{\boldsymbol{\beta}}'(\overline{V} + n_{12}^{-1}(B'B)^{-1})^{-1}\widehat{\boldsymbol{\beta}}$$

$\widehat{\sigma}^2$ is the standard maximum likelihood estimate of the variance of the combined sample,

$$\widehat{\boldsymbol{\beta}} = n_{12}^{-1}\sum_{j=1}^{2} n_j\widehat{\boldsymbol{\beta}}_j$$

and

$$rn_{12}\widehat{\sigma}^2 = \sum_{j=1,2} n_j r\widehat{\sigma}_j^2 + \frac{n_1 n_2}{n_{12}}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)'B'B(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)$$

We note that both models have a marginal likelihood which is a function of their hyperparameters and the four familiar statistics $\{\widehat{\boldsymbol{\beta}}_j, \widehat{\sigma}_j^2 : j = 1, 2\}$.

2.3. Bayesian MAP model selection

One popular method is Bayesian Maximum A Posteriori or MAP model selection (Bernardo and Smith, 1994). This simply chooses the model with the highest posterior probability. If the prior log odds for model M_t against model M_s are κ, then the distinct or separate vector model M_s is preferred to the combined vector model M_t when the posterior log odds are greater than κ. This occurs when $l_s(\boldsymbol{y}) - l_t(\boldsymbol{y}) > \kappa$ or, equivalently,

$$\Phi = \log(\overline{u} + \widehat{\boldsymbol{\beta}}_1'C_{11}\widehat{\boldsymbol{\beta}}_1 - 2\widehat{\boldsymbol{\beta}}_1'C_{12}\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\beta}}_2'C_{22}\widehat{\boldsymbol{\beta}}_2) - \sum_{j=1,2}\rho_j \log(u_j + \widehat{\boldsymbol{\beta}}_j'A_j\widehat{\boldsymbol{\beta}}_j) > \kappa'$$

where

$$\begin{array}{rclcrcl} A & = & (\overline{V} + \frac{1}{n_{12}}(B'B)^{-1})^{-1} & & A_j & = & (V_j + \frac{1}{n_j}(B'B)^{-1})^{-1} \\ C_{11} & = & \frac{n_1}{n_{12}}\left(\frac{n_1}{n_{12}}A + n_2B'B\right) & & \overline{u} & = & 2\overline{b} + r(n_1\widehat{\sigma}_1^2 + n_2\widehat{\sigma}_2^2) \\ C_{22} & = & \frac{n_2}{n_{12}}\left(\frac{n_2}{n_{12}}A + n_1B'B\right) & & u_j & = & 2b_j + rn_j\widehat{\sigma}_j^2 \\ C_{12} & = & \frac{n_1 n_2}{n_{12}}\left(\frac{1}{n_{12}}A + B'B\right) & & \rho_j & = & a_j^*/\overline{a}^* \end{array}$$

Note that the threshold

$$\kappa' = \left[2\kappa - \sum_{j=1,2} K(V_j, a_j, b_j, n_j) + K(\overline{V}, \overline{a}, \overline{b}, n_{12}) + 2(\overline{a} - a_1 - a_2)\log 2\right]/2\overline{a}^*$$

depends on the data only through (n_1, n_2) and the specified prior log odds κ between the two models. In principle, the prior parameter κ and hence κ' can take any value, so the

behaviour of this selection algorithm is formally explained simply through the geometry of the contours of the function Φ. For the remainder of the paper we will use the condensed notation $K(n)$ to denote $K(V, a, b, n)$.

The function Φ can be further simplified by introducing some new notation. We set $\mathbf{w}_j$ so that

$$\mathbf{w}_j' A_j \mathbf{w}_j = \mathbf{w}_j'(V_j + n_j^{-1}(B'B)^{-1})^{-1}\mathbf{w}_j = 1$$

Further, we define $\overline{z}_j = ||Q\widehat{\boldsymbol{\beta}}_j||$, where Q is any matrix satisfying $Q'Q = A_j$ and let $\lambda_1 = \mathbf{w}_1' C_{11}\mathbf{w}_1,\ \lambda_{12} = \mathbf{w}_1' C_{12}\mathbf{w}_2,\ \lambda_2 = \mathbf{w}_2' C_{22}\mathbf{w}_2$.

We then prefer M_s to M_t if and only if

$$\Phi = \log(\overline{u} + \lambda_1\overline{z}_1^2 - 2\lambda_{12}\overline{z}_1\overline{z}_2 + \lambda_2\overline{z}_2^2) - \sum_{j=1,2}\rho_j \log(u_j + \overline{z}_j^2) > \kappa'$$

Note that $(\overline{z}_1, \overline{z}_2)$ are the distances of the two profiles from zero, each scaled by a factor reflecting the deviation from zero we expected a priori under the separating model M_s. The statistics u_j depend on the data only through $\widehat{\sigma}_j^2$. The statistic $\overline{u}$ is a linear function of u_1 and u_2 and so is a linear function of the two corresponding sums of squares, and λ_j corresponds to the distance from zero expected for the profile $\widehat{\boldsymbol{\beta}}_j$ under M_t relative to that expected under M_s.

2.4. Using g-priors for conjugate clustering

Employing a general form of covariance matrix V demands that the space of prior hyper-parameters is very large. For simplicity, transparency and to ensure invariance to linear transformations of bases various authors (Chipman et al., 2001; Fernandez et al., 2001; Smith and Kohn, 1996; Zellner, 1986) have advocated the use of g-priors for prior covariance matrices.

In the given context, these priors would set $\overline{V}^{-1} = \overline{g}B'B$, $V_1^{-1} = g_1 B'B$, $V_2^{-1} = g_2 B'B$ for specified constants $(\overline{g}, g_1, g_2)$ associated with the combined cluster $\overline{c}$ and the smaller clusters c_1 and c_2. Here g is a measure of noise-to-signal so, in particular, the larger the value of g the greater the shrinkage of the expected posterior profile towards zero. Let $\mathbf{z}_j = (z_1^{(j)}, z_2^{(j)}, \ldots, z_p^{(j)})'$ with $j = 1, 2$ where

$$\mathbf{z}_j = \sqrt{\frac{n_j g_j}{g_j + n_j}} B\widehat{\boldsymbol{\beta}}_j$$

The vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ are the posterior expected profiles of the two clusters, normalised by their posterior variance. After some algebra it can be shown that the parameters of Φ then simplify to

$$\lambda_1 = \frac{(\overline{g} + n_2)(g_1 + n_1)}{(\overline{g} + n_{12})g_1}, \quad \lambda_2 = \frac{(\overline{g} + n_1)(g_2 + n_2)}{(\overline{g} + n_{12})g_2}, \quad \lambda_{12} = \lambda_{12}^0 \cos(\theta[\mathbf{z}_1, \mathbf{z}_2]),$$

where

$$\lambda_{12}^0 = \sqrt{\frac{n_1 n_2 (g_1 + n_1)(g_2 + n_2)}{(\overline{g} + n_{12})^2 g_1 g_2}}$$

The parameter $\theta[\mathbf{z}_1, \mathbf{z}_2]$ is the angle between vectors $(\mathbf{z}_1, \mathbf{z}_2)$ on a plane through zero containing the two rays $(\mathbf{0}, \mathbf{z}_1), (\mathbf{0}, \mathbf{z}_2)$. So, this is a measure of the difference in the scaled shapes of the two profiles.

A common choice of prior for model selection would be to set $g_1 = g_2 = \overline{g} = g$. This assumes that knowing the size, n, of a cluster would not affect the strength of our prior beliefs about the mean profile of a unit in that cluster. The prior information about each unit conditional on σ^2 is implicitly assumed to be based on exactly the same sources as other units in its cluster. We call this the *dependence* setting. Note that in this case

$$1 < \lambda_1 = \lambda_2 < 1 + \frac{\min\{n_1, n_2\}}{g}$$

An alternative protocol is sometimes applicable to, for example, gene expression data, where learning that a cluster of genes is large increases the chance that the cluster profile is close to zero: i.e. the cluster is not involved in regulation. A prior structure consistent with these beliefs — here called the *independence* model — assumes that the sources of information about the prior density of each single gene in a cluster are independent and of equal strength conditional on σ^2. This implies that $g_j = \breve{g} n_j$ and $\overline{g} = \breve{g} n_{12}$ so that

$$\lambda_1 = 1 + \frac{n_2}{\breve{g} n_{12}}, \quad \lambda_2 = 1 + \frac{n_1}{\breve{g} n_{12}}, \quad \lambda_{12}^0 = \frac{1}{\breve{g}} \sqrt{\frac{n_1 n_2}{n_{12}^2}}.$$

3. Using Bayes Factors to Select Between Many Partition Models

3.1. A typical example of conjugate Bayesian model selection

MAP model selection is used routinely in many tree and cluster models. In order to show how the performance characteristics of such selection can be linked to the study of the function Φ, we next review Bayesian model selection as it applies to the clustering algorithm in Heard et al. (2006). There, thousands of longitudinal profiles of genes are collected into a partition $C \in \mathbb{C}$ whose sets are the clusters $c \in C$. Microarrays measure the level of expression (a real number) for all of its genes over a sequence of times. In our running example there are 13 time points (Edwards et al., 2006).

The vector of profiles of the logged gene expressions, $\boldsymbol{Y}_c$, within each cluster are assumed to be exchangeable. $\boldsymbol{Y}^{(c)} = B\boldsymbol{\beta}_c + \varepsilon_c$, $\varepsilon_c \sim N(\mathbf{0}, \sigma_c^2 I_{rn_c})$, $\coprod \varepsilon_c | \boldsymbol{\beta}$, $\boldsymbol{\beta} = \text{vec}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N)$, where $\boldsymbol{Y}_c$ is a vector of length rn_c, where r is the length of the profile, n_c is the number of gene profiles in cluster c and N the number of sets in the partition C. Using analogous notation to that in section 2, we have that

$$\boldsymbol{Y}_i^{(c)} = B\boldsymbol{\beta}_c + \varepsilon_i^{(c)}$$

for $1 \leq i \leq n_c$, where $\varepsilon_i^{(c)} \sim N(\mathbf{0}, \sigma_c^2 I_r)$ and $\coprod_{i,c} \varepsilon_i^{(c)} | \boldsymbol{\beta}$, $c \in C$. The design matrix B is customised to the context. Thus a spline basis is employed in Heard et al. (2006), a Fourier basis is used in Anderson et al. (2006) and Edwards et al. (2006) and a wavelet basis is used in Ray and Mallick (2006). The profile vectors $\boldsymbol{\beta}_c$ and variances σ_c^2 of the different clusters $c \in C$ are all assumed to be mutually independent of each other and to follow the conjugate distributions given in section 2. So, in particular, each cluster has an associated multivariate t-distribution with log marginal likelihood $l_c(\boldsymbol{y})$. Furthermore, because of the assumed independencies between clusters in a given partition, the log marginal likelihood $l_C(\boldsymbol{y})$ of any partition C is simply the sum of the marginal likelihoods of its components:

$$l_C(\boldsymbol{y}) = \sum_{c \in C} l_c(\boldsymbol{y})$$

The log marginal likelihood of any partition can therefore be written down explicitly. Under MAP selection an optimal partition $C^* \in \mathbb{C}$ will be any partition such that, for all $C \in \mathbb{C}$,

$$l_{C^*}(\boldsymbol{y}) + \log \pi(C^*) \geq l_C(\boldsymbol{y}) + \log \pi(C)$$

where $\pi(C)$ is our prior probability that partition C generated the data.

3.2. Exchangeability and cohesions

To preserve certain exchangeability properties for partition models, the following four assumptions are commonly made (Barry and Hartigan, 1992; Quintana and Ingelias, 2003).

(a) The prior parameters (V_c, a_c, b_c) of cluster $c \in C$ depend on c but not C.
(b) The parameters (V_c, a_c, b_c) are a function of c only through n_c, the number of genes in c.
(c) The probabilities $\{\pi(C) : C \in \mathbb{C}\}$ satisfy

$$\pi(C) \propto \prod_{c \in C} \pi_c$$

where the proportionality constant is the sum of all these products of *cohesions*, π_c, over $C \in \mathbb{C}$.
(d) The probability π_c is allowed to depend on c only through its cardinality n_c.

We call prior beliefs for clustering *balanced* if they are consistent with these four assumptions. Previous studies (Anderson et al., 2006; Edwards et al., 2006; Heard et al., 2006) make a stronger assumption than (b) that (V_c, a_c, b_c) are not a function of n_c. The default choice of Heard et al. (2006) is balanced and sets cohesions so that $\pi_c = n_c!$. The appropriate choice of parametric form of a family of balanced priors - which determines the prior distribution of cardinalities of the vector of clusters in a given partition - is clearly highly dependent on the science and purpose underlying the statistical analysis. The default setting mentioned above tends to favour partitions with clusters of similar sizes, whilst Dirichlet priors tend to do reverse. However although this prior obviously influences which partition is optimal, all the instabilities we address in this paper apply whatever the choice of partition prior (see section 6). It is therefore possible to separate modeling issues associated with the three hyperparameters of each cluster from appropriate choices of balanced priors: an important issue but beyond the scope of this paper. Henceforth, when no confusion shall arise we will write $(V_{c_j}, a_{c_j}, b_{c_j}, n_{c_j})$ as (V_j, a_j, b_j, n_j), $j = 1, 2$.

We note that partition priors have been criticised because consistency is not preserved if exchangeability of units is demanded after deletion (McCullagh and Yang, 2006). However it is easily deduced that Ewens process priors which do have this consistency property lead to the same separation issues of hyperparameters from choice of partition prior parameter.

3.3. Model search

When the number of units partitioned is large (for example in Anderson et al. (2006) we clustered over $22,000$ genes), the partition space is huge. So, even being able to calculate the scores of single cluster partitions quickly is not enough to ensure that the scores of all the partitions in the vast partition space $\mathbb{C}$ can be evaluated. In practice it is therefore

often necessary to use an appropriate search algorithm to perform this optimisation task on a sensible subset of such partitions.

One useful feature of using $l_C(\boldsymbol{y})$ for selection is that the difference between the scores of two partitions identical outside a given set $\overline{c}$ will depend only on their relative scores over $\overline{c}$. We call partitions C^+ and C^- *adjacent* if the two partitions differ only on a set $\overline{c} \in C^+$ where $\overline{c} = c_1 \cup c_2$ with $c_1 \cap c_2 = \emptyset$, $c_1, c_2 \in C^-$ so that $\{c_1, c_2\}$ partition $\overline{c}$. Then

$$l_{C^-}(\boldsymbol{y}) = l_{C^+}(\boldsymbol{y}) - \Omega[C^-, C^+] - \log \pi(C^-) + \log \pi(C^+)$$

where

$$\Omega[C^-, C^+] = l_{c_1}(\boldsymbol{y}) + l_{c_2}(\boldsymbol{y}) - l_{\overline{c}}(\boldsymbol{y})$$

and $\pi(C^-)$, $\pi(C^+)$ are the prior probabilities of C^- and C^+ respectively. The comparison of adjacent partitions when using balanced priors is therefore especially straightforward and is utilised in many search algorithms used in this context. For example, the improvement presented by C^- (the model assuming the genes in $\overline{c}$ are in two different groups c_1 and c_2) over C^+ (the model assuming all genes in c are exchangeable) is measured by $\Phi - \kappa'(n_1, n_2)$ where

$$\kappa' = \frac{2\{\log \pi_{c_1}(n_1) + \log \pi_{c_2}(n_2) - \log \pi_{\overline{c}} n_{12}\} + K(n_1) + K(n_2) - K n_{12}}{2\overline{a}^*}$$

Note that κ' is a function of the two partitions only via a symmetry of the cardinalities (n_1, n_2) of the two potentially combined clusters. C^- has a higher posterior probability than C^+ if and only if $\Phi - \kappa'(n_1, n_2) > 0$. Thus, any search algorithm that moves only between adjacent partitions, either merging or splitting two clusters depending on whether the function Φ is large enough to instigate a split relative to a splitting penalty κ' (a function depending on cluster cardinalities within the relevant partitions but not on the data) is especially fast.

The most popular technique that uses adjacent moves to search a partition space is a greedy search algorithm called agglomerative hierarchical clustering (AHC) (Heard et al., 2006); a type of forward selection. This starts with each of the N gene profiles in N separate clusters with fixed values of the hyperparameters. A sequence of new partitions is then obtained by sequentially merging two clusters, thus decreasing the number of clusters by one. The two clusters chosen to be combined are the ones that increase the score (here the marginal likelihood of the partition) by the most (or reduce it by the least). Clusters are combined in this way until the trivial partition is reached, with one cluster containing all N genes. We have now calculated the marginal likelihood for a selection of N promising partitions containing 1 to N clusters. Finally we choose the partition in this sequence with the highest score: i.e. with the highest posterior probability over the partitions searched. Examples of other more elaborate search algorithms also using adjacent moves either in conjunction with a deterministic or stochastic search are given in Anderson et al. (2006) and, in a slightly different context, Chipman et al. (1998, 2002).

For the remainder of the paper we will study the geometry of $\Phi(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ as a function of the sufficient statistics $\{\widehat{\boldsymbol{\beta}}_j, \widehat{\sigma}_j^2 : j = 1, 2\}$ in order to understand the behaviour of MAP model selection methods using AHC. However we note that the problems we identify with the consequent model selection also apply to more sophisticated local search algorithms that allow clusters to be split as well as combined.

4. Bayesian Model Selection over Partitions

4.1. Three weaknesses of uncalibrated Bayesian model selection

Uncalibrated model selection based on Bayes factors like the one discussed above can fail for a number of reasons. Firstly, we have noted that the Bayes factor acts as an implicit real-valued score function over the different cluster partitions. There is thus an inevitable implicit trade-off between the closeness of the variances of the two potentially combined clusters and the closeness of their mean profiles. For this and other reasons, it is now well recognised that the chosen values of prior hyperparameters have a marked effect on the characteristics of Bayesian model selection, and their influence on inference cannot be expected to automatically fade away as the sample size increases. In fact, in section 7 we show how influential the selection of these hyperparameters is not only on the scale, but also on the *nature* of discrepancies that drive the selection. So there is great advantage to choose (whenever possible) prior values for hyperparameters not only so that the features of the selection algorithm match contextual knowledge, but also so that selection characteristics of the method are plausible a priori. As we discuss below, if this is not done, the properties of the induced selection algorithm can be absurd.

Secondly, as emphasised in Denison et al. (2002), the function Φ is not translation invariant. We demonstrate below that the optimal choice of partition is typically *critically* dependent on where we choose to set the prior mean vector of the profile — here we select zero. Hence unless, as in Edwards et al. (2006), there actually is a natural "preferred point", we cannot recommend the use of these methods. Henceforth we will assume, as is often the case in practice, that such a preferred point exists.

Thirdly the assumption of conjugacy is usually an expedience and there are at least two questionable consequences. First, the tails of the conjugate marginal likelihoods are inverse polynomials. Although this helpfully limits the number of small clusters, it also finds "optimal" partitions that often contain clusters that include outlying profiles. Second, these conjugate models imply that the prior mean and variance of the cluster profiles are quite highly dependent: for a careful discussion of this see O'Hagan and Forster (2004). One implication is that clusters observed to have an estimated profile very different from zero — our preferred point — will be allocated a high prior variance: a property which, if not recognised and adjusted for, can distort any search algorithm in ways discussed below.

4.2. Selection as a function of the magnitude of the mean profile

From the comments above we might suspect model selection to be disrupted by outliers. Consider the effects of increasing the magnitude of a cluster profile away from zero whilst holding all other statistics fixed. Fix $\overline{z}_2$, $\mathbf{w}_j$, $\widehat{\sigma}_j^2$ and n_j for $j = 1, 2..$ Then, provided $0 < \rho_1 < 1$,

$$\lim_{|z_1| \to \infty} \Phi(\overline{z}_1, \overline{z}_2) = \infty$$

Thus whatever the values of prior hyperparameters, as we increase the magnitude $\overline{z}_1$ of the profile of the first cluster c_1 (provided $\overline{z}_1$ is large enough) our model will prefer to keep clusters c_1 and c_2 separate, as we might hope.

However, if *two* outlying clusters (c_1, c_2) both have profiles $(\overline{z}_1, \overline{z}_2)$ far from zero then model selection can start to display strange properties. If $\overline{z}_2 = l\overline{z}_1^k$, l is fixed and $|\overline{z}_1| \to +\infty$, then $\Phi(\overline{z}_1, \overline{z}_2)$ diverges to $-\infty$ if $\rho_1 + k\rho_2 > 1$ and diverges to $+\infty$ if $\rho_1 + k\rho_2 < 1$. For

example, Heard et al. (2006) recommend setting $\overline{a} = a_1 = a_2$. This implies that

$$\rho_1 + \rho_2 = \frac{4\overline{a} + n}{2\overline{a} + n} > 1$$

where $n = rn_{12}$ is the total number of observations associated with the two groups. In this case, by simultaneously increasing the magnitude of the two cluster profiles by the same amount $z_1 = z_2$, we will eventually reach a magnitude where two clusters are combined *irrespective* of how different the shapes of those clusters are: definitely not what we want to happen. This occurs because, when combined into one cluster, these two outliers become one outlier and, a priori, one outlier is assumed more probable than two.

Thus two clusters whose expression profiles are far from zero — and hence possibly biologically significant — will be combined in preference to any other pair: even clusters whose statistics are identical! The reason this unfortunate property is relatively rare in practice is that studies such as Heard et al. (2006) happen to suggest the use of a small value of $\overline{a}$. Therefore profiles have to be very different from zero before this phenomenon can be realised. However, this still happens even at the recommended settings of the parameters. In figure 1 we can see that genes with completely different profiles have been attracted into a cluster under an optimal MAP partition found under an AHC search. Note that when this phenomenon occurs early in an AHC search, the combined cluster can largely cancel out and then has the signature of the large variance cluster: something we term a junk cluster in Anderson et al. (2006). When such a cluster is formed under AHC it tends to act as an attractor to yet more disparate and biologically interesting clusters resulting in a cluster like the one depicted in figure 2.

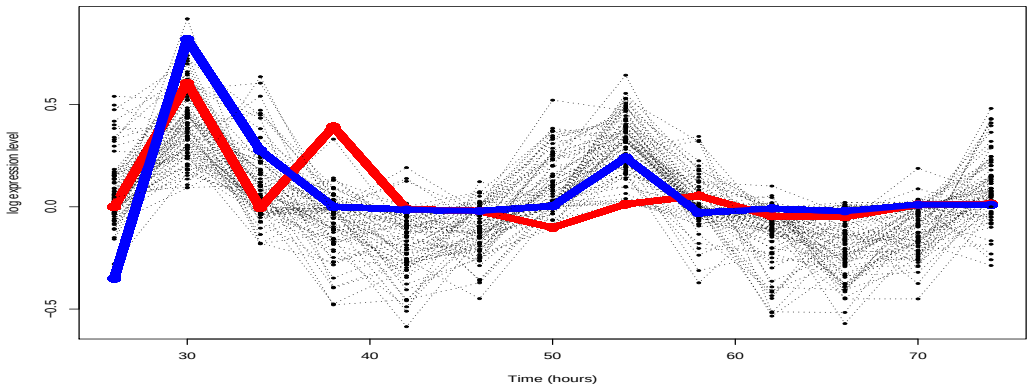

Fig. 1. A cluster of 81 gene expression profiles from a early stage of the clustering performed in Edwards et al. (2006) using the default hyperparameter settings. The two highlighted genes are clearly outliers that do not belong in this cluster. This is a result of the AHC and the default settings.

If we differ from Heard et al. (2006) and choose a prior with $\rho_1 + \rho_2 < 1$, then $\Phi(\overline{z}_1, \overline{z}_2)$ $\to \infty$ as $|\overline{z}_1| \to \infty$. This gives rise to an even more problematic property. Whatever our settings of prior hyperparameters, two profiles sufficiently far from zero will always be put in separate clusters even when $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_2$, $\widehat{\sigma}_1^2 = \widehat{\sigma}_2^2$ and $n_1 = n_2$, i.e., even when these two clusters are identical in all respects! Note that the position of the prior mean (here the zero setting) is central to determining which profiles are outlying in the sense above.

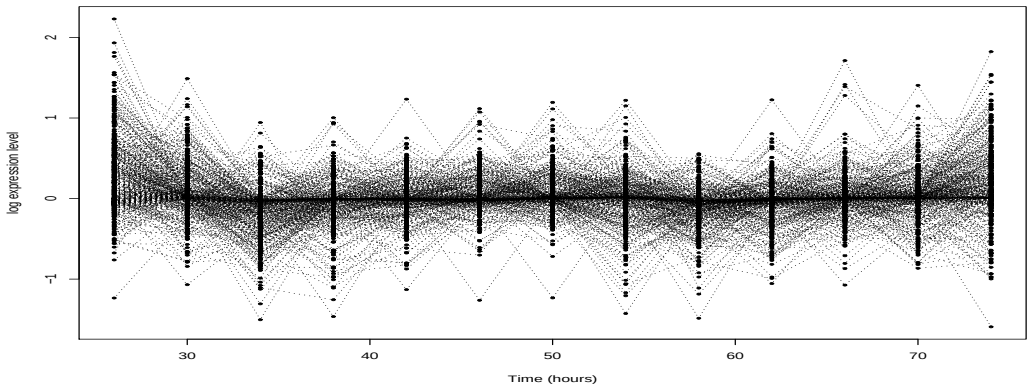

Fig. 2. A cluster of 453 gene expression profiles from the same partition as figure 1. This so-called junk cluster is a by-product of AHC and contains a broad variety of profile shapes. Note that in this context log expressions outside $[-0.5, 0.5]$ are considered to be potentially of biological interest.

The *only* case when the associated limit stays finite is when $\rho_1 + \rho_2 = 1$. Unless we set the hyperparameters to ensure this, on observing profiles far from zero the implications of the prior are unlikely to be faithful to contextual beliefs. Therefore, the Bayesian clustering algorithm will be prone to perform inappropriately, and combine profiles it was never meant to.

4.3. Models with $\rho_1 + \rho_2 = 1$

By setting hyperparameters so that $\rho_1 + \rho_2 = 1$ the characteristics of the resulting merging criterion are much more compelling. The demand that $\rho_1 + \rho_2 = 1$ is satisfied provided that the hyperparameters (a_1, a_2) of two clusters in a partition and the hyperparameter $\overline{a}$ of the combined cluster in an adjacent partition satisfy

$$a_1 + a_2 = \overline{a}$$

For balanced priors, this implies that we set the corresponding hyperparameter $a_c = \breve{a}n_c$, where n_c is the number of profiles in c rather than require a_c to be independent of cluster size as is the case in Heard et al. (2006). Our suggestion would make the prior coefficient of variation of the precision of a cluster proportional to $n_c^{1/2}$. For example, in the context of gene clustering this would mean that 'genuine' clusters containing large numbers of gene profiles are expected to have smaller associated coefficients of variation in their precision. Thus we are a priori less certain about the value of the variance of big clusters: not an unreasonable assumption in this context. Note that under this setting $\rho_j = n_j n_{12}^{-1}$, $j = 1, 2$.

5. Bayes Factors and Measures of Separation

Under balanced priors, each cluster c in a partition has a set of sufficient statistics $\mathbf{x}(c) = (n_c^{-1}\widehat{\boldsymbol{\beta}}_c, \widehat{\sigma}_c^2, n_c)$. Let $\kappa'' = \min \Phi$. Then, it is common (Denison et al., 2002) to interpret the function $\Delta = \Phi - \kappa''$ as a measure of the separation between the combined clusters c_1 and c_2 in two adjacent partitions identical except on $c_1 \cup c_2$. We have seen above that this

interpretation may well not be correct. Whenever $\rho_1 + \rho_2 \neq 1$, two clusters c_1 and c_2 with identical sufficient statistics can be arbitrarily more separated — i.e. have an arbitrarily higher value of Δ — than two clusters that have very different sufficient statistics. In particular under *any* search over the partition space, it is quite possible for two clusters with widely differing profiles to be combined in preference to two clusters with identical $\{\widehat{\boldsymbol{\beta}}_j, \widehat{\sigma}_j^2 : j = 1, 2\}$.

Although this phenomenon is much more dramatic when $\rho_1 + \rho_2 \neq 1$, the problem can still remain even when hyperparameters are set so as to ensure $\rho_1 + \rho_2 = 1$. In this section we investigate to what extent, with appropriate parameter settings, $\Delta(\mathbf{x}(c_1), \mathbf{x}(c_2))$ can be interpreted as a measure of separation between the clusters c_1 and c_2.

If $\Psi(\mathbf{x}(c_1), \mathbf{x}(c_2)) = f_1(\Delta(\mathbf{x}(c_1), \mathbf{x}(c_2))) + f_2(n_1, n_2)$ where f_1 is some strictly increasing function of $\Delta(\mathbf{x}(c_1), \mathbf{x}(c_2))$ and f_2 is an arbitrary penalty function on the size of clusters then a property that would normally be required of a separation measure is that for any two clusters c_1 and c_2 that have identical characteristics, so that $\mathbf{x}(c_1) = \mathbf{x}(c_2)$, we have

$$\Delta(\mathbf{x}(c_1), \mathbf{x}(c_2)) = 0 \tag{1}$$

At this point it is convenient to re-parametrise Φ. Let $d = (\overline{z}_1^2 + \overline{z}_1^2)\overline{u}^{-1}$ represent a normalised squared distance from zero of the two clusters, define $\alpha_j = d^{-1}\overline{u}^{-1}\overline{z}_j^2$, $j = 1, 2$ to be the corresponding relative squared distance from zero of the two clusters (so that in particular $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 + \alpha_2 = 1$) and let $v_j = u_j\overline{u}^{-1}$, $j = 1, 2$ be approximately the relative sums of squares of the two clusters. Then

$$\gamma = \lambda_1\alpha_1 - 2\lambda_{12}\sqrt{\alpha_1\alpha_2} + \lambda_2\alpha_2$$

and

$$\Phi = \log(1 + \gamma d) - \rho_1 \log(v_1 + \alpha_1 d) - \rho_2 \log(v_2 + \alpha_2 d)$$

DEFINITION 1. *Define* $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2))$ *as* homogeneous *if, whenever* $\mathbf{x}(c_1) = \mathbf{x}(c_2)$, $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2)) = \Phi_0$ *is a function of* (n_1, n_2) *alone.*

Under the family of separations above, a necessary and sufficient condition for $\Psi(\mathbf{x}(c_1), \mathbf{x}(c_2))$ to satisfy the property leading to equation (1) is that $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2))$ is homogeneous.

THEOREM 1. *If* $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2))$ *is homogeneous and a g-prior is employed then for any two identical clusters* c_1 *and* c_2 *such that* $\bar{n} = 2n_1$,

$$\bar{a} = 2a_1, \quad \bar{b} = 2b_1 \quad \textit{and} \quad \bar{g} = 2g_1.$$

Furthermore, if these three conditions above hold, then $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2))$ *will be homogeneous.*

See Appendix A.1 for the proof. Note that the standard way of assigning a prior to a conjugate model is not homogeneous and so falls at the first hurdle. However there is an obvious family of conjugate Bayesian models which is homogeneous.

COROLLARY 1. *The* proportional model *which sets* $a_c = \breve{a}n_c$, $b_c = \breve{b}n_c$ *and* $g_c = \breve{g}n_c$ *for some values* $\breve{a}, \breve{b}, \breve{g} > 0$ *is homogeneous.*

For the proportional model, $\rho_j = n_j n_{12}^{-1}$, $u_j = (2\breve{b} + r\widehat{\sigma}_j^2)n_j$ and $\overline{u} = u_1 + u_2$ so that $v_1 + v_2 = 1$. Furthermore, let the value of γ when two profiles are identically oriented (so that $\theta[\mathbf{z}_1, \mathbf{z}_2] = 0$) be γ_0. Then under the proportional model

$$\gamma_0 = 1 + (\sqrt{\rho_2\alpha_1} - \sqrt{\rho_1\alpha_2})^2 \breve{g}^{-1}$$

We can now derive some important properties of proportional clustering (see Appendix A.2 for the proof).

THEOREM 2. *Under proportional clustering, for all possible values of* $\mathbf{x}_1(c_1), \mathbf{x}_2(c_2)$,

$$\Phi(\mathbf{x}_1(c_1), \mathbf{x}_2(c_2)) \geq I(\rho)$$

where $\rho_j = n_j n_{12}^{-1}, j = 1, 2$, *and* $I(\boldsymbol{\rho}) = -\sum_{j=1,2} \rho_j \log \rho_j$.

COROLLARY 2. *For any fixed (unordered) pair* $\mathbf{n} = (n_1, n_2)$

$$\Delta_{\mathbf{n}}(\mathbf{x}(c_1), \mathbf{x}(c_2)) = \Phi(\mathbf{x}_1(c_1), \mathbf{x}_2(c_2)) + \kappa''(\mathbf{n})$$

where $\kappa''(\mathbf{n}) = -I(\boldsymbol{\rho}) - \kappa'$ *is a separation measure. That is,*

(a) For all pairs $(\mathbf{x}(c_1), \mathbf{x}(c_2))$

$$\Delta_{\mathbf{n}}(\mathbf{x}(c_1), \mathbf{x}(c_2)) \geq 0$$

with equality if and only if $\mathbf{x}(c_1) = \mathbf{x}(c_2)$

(b) For all pairs $(\mathbf{x}(c_1), \mathbf{x}(c_2))$

$$\Delta_{\mathbf{n}}(\mathbf{x}(c_1), \mathbf{x}(c_2)) = \Delta_{\mathbf{n}}(\mathbf{x}(c_2), \mathbf{x}(c_1))$$

See Appendix A.3 for the proof of Corollary 2. So a sufficient and almost necessary condition for MAP selection to behave in a way that combines clusters in partitions with "close" statistics is that the hyperparameters are set as a proportional model. For most other settings, and in particular those advocated by other authors as defaults, this is not the case. It is interesting to note that to ensure consistency in different contexts various authors have suggested introducing a dependency of the parameter g on sample size. However, this suggested dependency demands that the prior variance of the proportional model decreases in the cluster size n whereas here it increases. This is not too disturbing for our applications. The natural type of consistency we might require here is associated with the length of profile — a function of the experimental design — not the number of genes of certain types which is determined by the technology of the gene chip and thus fixed. Note that with the hyperparameter settings recommended here, consistency is automatic under increasing profile length.

6. Comparison for Two Simple Simulation Studies

In order to illustrate the characteristics of cluster inference under the conventional settings of the hyperparameters as described by Heard et al. (2006) and our proportional setting, we have simulated from scenarios where the desired characteristics of the clustering algorithm are fairly transparent.

6.1. Outliers and junk clusters

First consider clustering just 7 points (profiles of length 1) simulated from 3 clusters of sizes $n_1 = 2$, $n_2 = 4$ and $n_3 = 1$. The two points in the first cluster are drawn from a distribution with a large negative mean expression $-s$, the points in the second cluster drawn have zero mean expression and the point in the fourth cluster has large expression s. So the means of the 7 points cluster into the partition $A = \{(-s, -s), (0, 0, 0, 0), (s)\}$. In our notation

$$\boldsymbol{Y}^{(j)} = B\boldsymbol{\beta}_j + \varepsilon^{(j)}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3) = (-s, 0, s)'$, $B = 1$ and $\varepsilon^{(j)} \sim N(0, 0.05)$ for $j = 1, 2, 3$ and we set $s = 1,000$.

Note that whilst the undesirable partition $B = \{(-s, -s, s), (0, 0, 0, 0)\}$ appears as candidate partition in both methods, as is typical, it appears earlier under the conventional settings than the proportional settings.

To compare the proportional scaling method with that of Heard et al. (2006) for simplicity we have subsequently set $\breve{g} = g$, $\breve{a} = a$, $\breve{b} = b$, so that the algorithms exactly correspond at the beginning. For comparability we use the same default prior as Heard et al. (2006) over the partition space.

We now compare the performance of the clustering algorithm by Heard et al. (2006) for different values of the prior parameters to ours in figure 4. A typical dendrogram of the combination under the conventional setting and default partition priors is given in figure 3 together with another dendrogram which is often produced by the algorithm by Heard et al. (2006). Note that in the second dendrogram in figure 3, as anticipated in section 4.2, the first cluster combines the three outliers at an early step, a combination that under AHC can never be retrieved. Such unhelpful properties obviously depend on the setting of the hyperparameters.

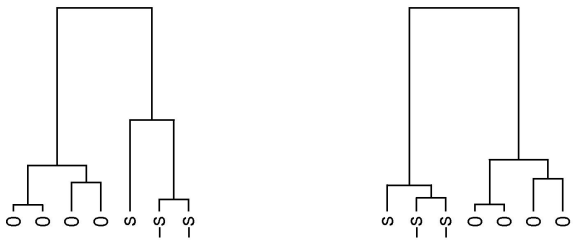

Fig. 3. Two of the dendrograms produced by AHC using the algorithm with proportional parameters and the algorithm by Heard et al. (2006).

Obviously the precise combination reflected in such dendrograms depends on how the values of the hyperparameters are chosen. So in figure 4 we have determined which values of the simulated data sets correctly identified the true simulated partition for the conventional and our settings of hyperparameter (identified as above) and default choice of prior by Heard et al. (2006) over the partition space. Notice that our method appears much more stable to misspecification of these three hyperparameters. The values used are $g = \breve{g}^{-1} \in [1, 10]$, $a = \breve{a} \in [0.01, 1]$ and $b = \breve{b} \in [0.01, 1]$.

Finally in figure 5 we compare the number of times during the simulations the partitions A, B, the 'large variance' partition $C = \{(-s, -s, 0, 0, 0, 0, s)\}$ and all other partitions D are chosen as optimal. Notice that the broad effect here is for the vast proportion of

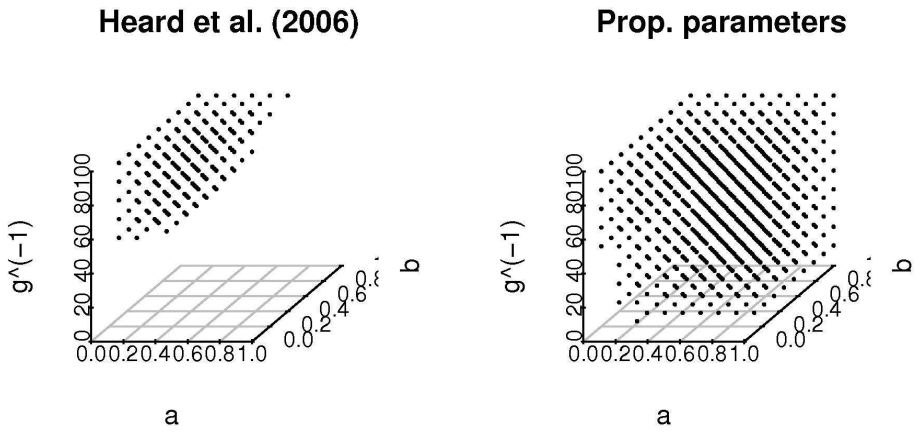

Fig. 4. Result of the clustering algorithm by Heard et al. (2006) and our algorithm for different values of the prior parameters. Each dot corresponds to a combination of values of the prior parameters which generated the desired partition A of our dataset.

partitions misclassified as B under conventional clustering to be properly clustered as A under proportional clustering.

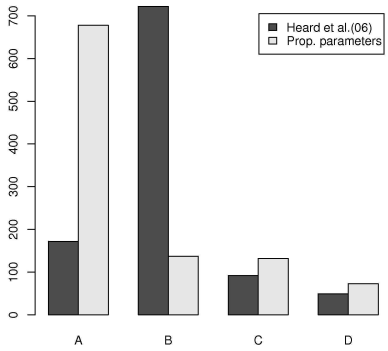

Fig. 5. When different prior parameters are used, the algorithms produce different partitions of our dataset. The plot above shows the counts of each partition produced by the algorithm by Heard et al. (2006) and the algorithm with proportional parameters.

6.2. Merging of complementary profiles

A property of a clustering algorithm we would like to avoid is one where two complementary profiles (i.e. two profiles where one is approximately the negative of the other, each with high expression) are combined into a single large variance approximately zero mean cluster. In our second simulation we therefore created such a scenario. Typically for higher dimensional problems we introduce further tuning parameters on the prior over the partition. However the parameters have no effect on the combination of clusters when they are all of the same cardinality - as they are at the beginning of the AHC algorithm. We can therefore compare our algorithm fairly with the conventional one with default prior if we focus on the behaviour

of the algorithm on the *first combination* of the AHC.

Thus consider the dataset formed by the following three clusters

$$\begin{aligned} \boldsymbol{Y}_k^{(1)} &= B\boldsymbol{\beta}^{(1)} + \varepsilon_k \qquad k = 1 \\ \boldsymbol{Y}_k^{(2)} &= B\boldsymbol{\beta}^{(2)} + \varepsilon_k \qquad k = 2, \ldots, 6 \\ \boldsymbol{Y}_k^{(3)} &= -B\boldsymbol{\beta}^{(1)} + \varepsilon_k \quad k = 7 \end{aligned} \tag{2}$$

where $\varepsilon_k \sim N(0, 1)$ for $k = 1, \ldots, 7$ and B is the Fourier design matrix as in Anderson et al. (2006).

Following the notation and vocabulary of the running example as in Anderson et al. (2006), our dataset, drawn in figure 6, has two complementary gene profiles and 5 gene profiles close to zero. Note that genes in cluster 1 and 3 have opposite complementary profiles, so it is critical not to combine genes from these two different profiles into a single cluster. Cluster 2 represents a set of unresponsive genes with a zero mean profile.

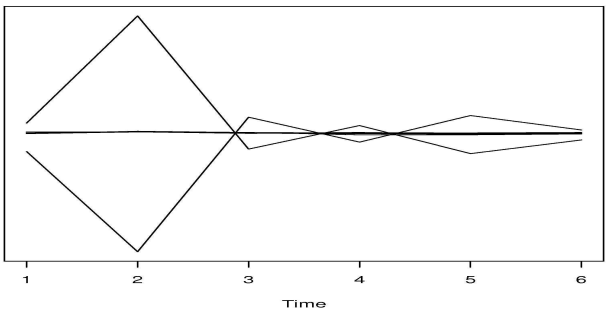

Fig. 6. Data simulated as in model (2). The amplitude of the curves is variable and depends on the value of σ^2.

The worst case scenario happens when the observations in clusters 1 and 3 are combined together at the first iteration of the algorithm. When this happens under AHC the algorithm can never identify the desired partition, which therefore will not be identified no matter which priors we are using on the partitions. Consider the results in 1. The prior parameters used were $g = \breve{g} \in [1, 1000]$, $a = \breve{a} \in [0.01, 1]$ and $b = \breve{b} \in [0.01, 1]$. Again it is easy to see how our new settings improve on the original in this circumstance, particularly when expressions are large.

7. Separation of Models: Separation of Statistics

7.1. Some useful parameters

Although we have found a separation measure corresponding to Bayesian selection, it remains to demonstrate that this induced measure is largely consistent with a separation measure with which we would be content predictively. We therefore next examine how the function $\Phi = \Delta^{(1)} + \Delta^{(2)} + I(\boldsymbol{\rho})$ where

$$\begin{aligned} \Delta^{(1)} &= \log(1 + \gamma d) - \log(1 + d) \\ \Delta^{(2)} &= \log(1 + d) - \rho_1 \log(v_1 + \alpha_1 d) - \rho_2 \log(v_2 + \alpha_2 d) - I(\boldsymbol{\rho}), \end{aligned}$$

Table 1. The table shows the number of times (out of 432) that genes in cluster 1 and 3 are combined together at the first step of the algorithm by Heard et al. (2006) and the algorithm with proportional parameters as the amplitude of the curves increases.

σ^2	Heard et al. (2006)	Our algorithm
10	3	0
100	8	0
1,000	32	0
10,000	62	0
100,000	95	0
1,000,000	102	0

compares adjacent partitions for the proportional model as a function of the sufficient statistics of two profiles. This allows us both to confirm that the characteristics of the induced separation measure are largely desirable and guides us to settings of prior hyperparameters that ensure plausible predictive implications. Because we need to acknowledge that the Bayes factor clustering has an intrinsic structure that selects as a function of (n_1, n_2), in this section we will assume the cardinalities (n_1, n_2) of two candidate clusters — and hence (ρ_1, ρ_2) — are fixed. There are four statistics that are central to the combination rule: d, v_1 (defined above), η and ζ^2 (defined below).

(a) The statistic $\eta = \sqrt{\alpha_1\alpha_2}(1 - \cos(\theta[\mathbf{z}_1, \mathbf{z}_2])$ is a weighted measure of the angle between the two cluster profiles, taking a value of zero when the posterior expected profiles are proportional to one another and its maximum value when the profiles are proportional to one another but of opposite sign: as would be the case whenever one gene is up-regulated the other is down-regulated. Thus η is a measure of the *dissimilarity in orientation* of the two profiles.

(b) A measure of the *differences in overall magnitudes* of squared differences in distance from zero relative to that expected under the given cluster size under the prior, ζ^2, satisfies

$$0 \leq \zeta^2 = (\sqrt{\rho_2\alpha_1} - \sqrt{\rho_1\alpha_2})^2 \leq \max\{\rho_1, \rho_2\}$$

Note that $\zeta = \sin(\sin^{-1}(\sqrt{\alpha_1}) - \sin^{-1}(\sqrt{\rho_1}))$ and so, for fixed ρ_1, ζ is an invertible function of α_1.

Now

$$\Delta_{\mathbf{n}}^{(1)}(\gamma(\eta, \zeta), d) = \log(1 + (\gamma - 1)(1 + d^{-1})^{-1})$$

where

$$1 \leq \gamma = 1 + (\zeta^2 + 2\eta\sqrt{\rho_1\rho_2})\breve{g}^{-1} \tag{3}$$

and

$$\Delta_{\mathbf{n}}^{(2)}(v_1, \zeta, d) = \log(1 + d) - \sum_{j=1,2} \rho_j \log(v_j + \alpha_j d) - I(\boldsymbol{\rho})$$

Note that $\Delta_{\mathbf{n}}^{(1)}$ is a function only of (η, ζ, d) and ignores (v_1, v_2) whilst $\Delta_{\mathbf{n}}^{(2)}$ is a function of relative variances, relative size ζ expressed as a function of α_1, and combined size d and also ignores η.

7.2. Angular separation, $\Delta_{\mathbf{n}}^{(1)}$

The following results are straightforward to verify. $\Delta_{\mathbf{n}}^{(1)}(\gamma(\eta,\zeta),d)$ is strictly increasing in η,ζ and d with

$$\begin{aligned}\lim_{\eta\to 0}\Delta_{\mathbf{n}}^{(1)} &= \log(1+\zeta^2(1+d^{-1})^{-1}\breve{g}^{-1})\geq 0\\ \sup_{d}\Delta_{\mathbf{n}}^{(1)} &= \log(1+(\max\{\rho_1,\rho_2\}+2\sqrt{\rho_1\rho_2})\breve{g}^{-1})\\ \lim_{d\to 0}\Delta_{\mathbf{n}}^{(1)} &= 0\end{aligned}$$

Note that this function is bounded. Its contribution to the selection depends on the prior noise-to-signal parameter $\breve{g}$. Thus if $\breve{g}$ is large, so that observational error is assumed to dominate the signal, the contribution of this term to the selection is negligible. On the other hand, if $\breve{g}$ is small, the difference in orientation between the two profiles is smaller. The oriented distance is weighted in this function by $\sqrt{\alpha_1\alpha_2}$ so that similar length profiles and similar orientations are made less prone to combination than similar length profiles with different orientations, whilst the term $\sqrt{\rho_1\rho_2}$ ensures this penalty only bites for clusters of comparable magnitude. The further apart the posterior expected profiles of the clusters are from zero, the more inclined we are to keep these separate.

7.3. Relative distance/variance separation, $\Delta_{\mathbf{n}}^{(2)}$

The second component $\Delta_{\mathbf{n}}^{(2)}(v_1,\alpha_1,d)=\log(1+d)-\rho_1\log(v_1+\alpha_1 d)-\rho_2\log(v_2+\alpha_2 d)-I(\boldsymbol{\rho})$ is a function of the relative sums of squares and scaled relative distances from zero but not $\breve{g}$. unlike $\Delta_{\mathbf{n}}^{(1)}(\gamma,d)$ it is unbounded above and, depending on the distance d of the two clusters from zero, can heavily penalise the combination of clusters with relatively very different associated estimated variances or different scaled lengths from the origin. Thus, for example,

$$\begin{aligned}\lim_{d\to 0}\Delta_{\mathbf{n}}^{(2)}(v_1,\alpha_1,d) &= \rho_1\log\frac{\rho_1}{v_1}+\rho_2\log\frac{\rho_2}{v_2}\\ \lim_{d\to\infty}\Delta_{\mathbf{n}}^{(2)}(v_1,\alpha_1,d) &= \rho_1\log\frac{\rho_1}{\alpha_1}+\rho_2\log\frac{\rho_2}{\alpha_2}\\ \lim_{v_1=\alpha_1\to 0}\Delta_{\mathbf{n}}^{(2)}(v_1,\alpha_1,d) &= \lim_{v_1=\alpha_1\to 1}\Delta_{\mathbf{n}}(v_1,\alpha_1,d)=\infty\end{aligned}$$

So when d is small and the two profiles are close to zero, $\Delta_{\mathbf{n}}$ acts as a penalty mainly for divergent estimates of variance, taking close to its minimum value whenever $\widehat{\sigma}_1^2=\widehat{\sigma}_2^2$. However when d is very large it penalises almost entirely on the basis of the difference in distance of the two clusters from the origin and ignores any divergence in their estimated variances. So under the AHC algorithm, when all the clusters in a partition are about the same cardinality, the Bayes factor algorithm will still tend to encourage the combination of clusters far from zero with the same orientation and distance from zero even when their estimated variances are very different.

It is easily checked that the stationary points of $\Delta_{\mathbf{n}}^{(2)}$ are solutions of $P+Qd=0$ where

$$\begin{aligned}P &= v_1v_2-\rho_1\alpha_1v_2-\rho_2\alpha_2v_1=v_1v_2\left(1-\rho_1\alpha_1v_1^{-1}-\rho_2\alpha_2v_2^{-1}\right)\\ Q &= \rho_1v_1\alpha_2+\rho_2v_2\alpha_1-\alpha_1\alpha_2=-\alpha_1\alpha_2\left(1-\rho_1v_1\alpha_1^{-1}-\rho_2v_2\alpha_2^{-1}\right)\end{aligned}$$

If we set $\alpha_1 = 0.5(1-\omega)$ and $v_1 = 0.5(1-\varepsilon)$ then it follows that $\Delta^{(2)}_{\mathbf{n}}$ will have a non-zero feasible stationary point in d if and only if ω and ε have different signs. Otherwise, $\Delta^{(2)}_{\mathbf{n}}$ is monotonic in d. Note that when $\rho_1 = \rho_2 = 0.5$, $\Delta^{(2)}_{\mathbf{n}}$ has a stationary point $d^* = \frac{v_1 - v_2}{\alpha_2 - \alpha_1}$ if and only if $\frac{v_1 - v_2}{\alpha_2 - \alpha_1} \geq 0$. So in this case the stationary point is a minimum. When $d = 0$, $\Delta^{(2)}_{\mathbf{n}} = -0.5\log(1-\varepsilon^2)$. As $d \to \infty$, $\Delta^{(2)}_{\mathbf{n}} \to -0.5\log(1-\omega^2)$ so $d = 0$ and $d = \infty$ are the two local maxima of this function.

Whatever the value of ρ_1, $v_1 = \alpha_1 \Rightarrow P = Q = 0$ so that $\Delta^{(2)}$ is not a function of d and takes the value

$$\Delta^{(2)}_{\mathbf{n}}(v_1, \zeta, d) = \sum_{j=1,2} \rho_j \log \frac{\rho_j}{v_j}$$

Thus the characteristics of the induced separation measure of the proportional model seem eminently desirable, with the caveat that the conjugacy encourages outlying clusters with similar profiles but different variances to occasionally be combined when the two clusters are far from zero. However it is easily verified that when clusters are about the same cardinality, so that $\rho_1 \simeq \rho_2$, and ε, ω are small then this dependence on d is insignificant.

7.4. Combined separation

Finally, to appreciate how the relative magnitude of clusters is traded with its distance from zero, think of Φ as a function of d only and fix α_1. We can then calculate that its stationary points satisfy the equation

$$P' + Q'd = 0$$

where

$$\begin{aligned} P' &= \gamma v_1 v_2 - v_2\alpha_1\rho_1 - v_1\alpha_2\rho_2 \\ Q' &= \gamma(\alpha_1 v_2 + \alpha_2 v_1) - \alpha_1\rho_1(\gamma v_2 + \alpha_2) - \alpha_2\rho_2(\gamma v_1 + \alpha_1) \end{aligned}$$

Thus, in particular, there is at most one stationary point of Φ for $d \in \mathbb{R}_{>0}$. This will be located at $d^* = -P'Q'^{-1}$ provided $P'Q'^{-1} < 0$. Note that when $\breve{g}$ is small, except when $|v_1 v_2^{-1}|$ is $\Delta^{(2)}_{\mathbf{n}}$ will dominate this expression so that Φ is simply increasing in d.

Note that the geometry of Φ is simple because $\rho_1 + \rho_2 = 1$. When $\rho_1 + \rho_2 \neq 1$ it is easily verified that the stationary points lie on a quadratic, giving rise to a much richer geometry in Φ. This is the algebraic reason for much of the strangeness of the induced selection. This phenomenon is illustrated in the central column of figure 7 where we graph Φ for two clusters with identical orientation and cardinality for various settings of the hyperparameters. The dependence of Φ on d is only significant when Φ takes large values. In this case the clusters will usually be kept separate for other reasons anyway. Dependence on the relative distance from zero of the two clusters only occurs when d is of moderate magnitude. Furthermore the discrepancy in relative variances is only significant when their ratio is substantially different from one and then only when d is quite far from zero.

The left hand column of figure 7 illustrates the phenomenon discussed in section 4.2 that when $\rho_1 + \rho_2 < 1$, clusters become increasingly large the further d is from zero: Φ eventually becoming very large regardless of how close the pair of cluster statistics are. On the other hand the right hand column (when $\rho_1 + \rho_2 > 1$) shows as d becomes large, the two clusters will become close regardless of the value of the other statistics. This illustrates why Bayes Factor selection can be badly behaved unless the hyperparameters are chosen carefully.

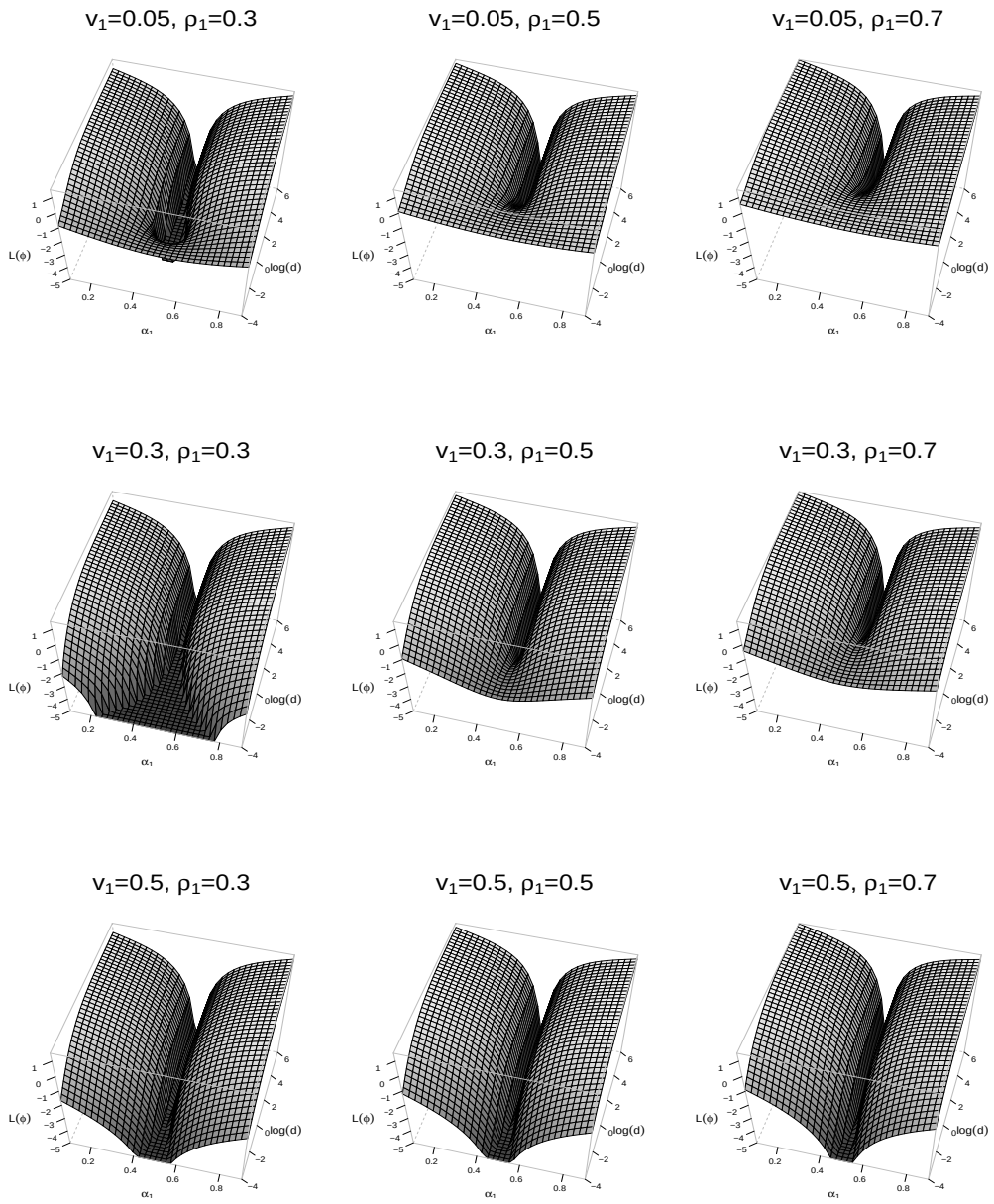

Fig. 7. A plot of $L(\Phi) = \log(\max\{\Phi - \log(2), \exp(-5)\})$. The magnitude of this measure represents the inclination of two clusters to merge, d is the mutual distance from zero of the two clusters, α_1 is the distance from zero of cluster 1 relative to this mutual distance, and v_1 is the relative variance of the profile of cluster 1 relative to that of the combined cluster for the proportional model. The central column shows the recommended setting when $\rho_1+\rho_2 = 1$ whilst the left and right hand columns show $L(\Phi)$ when $\rho_1 + \rho_2 < 1$ and $\rho_1 + \rho_2 > 1$ respectively. Here, $v_1 = \{0.05, 0.3, 0.5\}$, $\rho_1 = \{0.3, 0.5, 0.7\}$ and $\alpha_1 \in [0.1, 0.9]$, $\log(d) \in [-4, 6]$. We choose $\rho_1 = \rho_2$, $\eta = 0$ (since the two clusters are identically oriented) and $\breve{g}^{-1} = 100$ so that equation (3) implies that $\gamma = 1 + 100\rho_1(\sqrt{\alpha_1} - \sqrt{1-\alpha_1})^2$. Note that $\alpha_1 + \alpha_2 = 1$ and $v_1 + v_2 = 1$ are always satisfied. Equation (3) has singularities at $v_1, \alpha_1 = \{0, 1\}$. All nine plots have the same axes.

7.5. Characteristics of non-proportional models

We have seen that when $\rho_1 + \rho_2 \neq 1$ the function Φ does not behave anything like a separation measure on the other components. On the other hand, when $\rho_1 + \rho_2 = 1$ the characteristics of Φ can still sometimes act as an approximate separation measure. Writing $v_j = u_j \overline{u}^{-1}$ for $j = 1, 2$ and $d = z\overline{u}^{-1}$

$$\Phi(d, \gamma, v_1, v_2, \alpha_1, \rho_1) = \log(1 + \gamma d) - \rho_1 \log(v_1 + \alpha_1 d) - (1 - \rho_1) \log(v_2 + (1 - \alpha_1)d)$$

so that

$$\lim_{d \to \infty} \Phi(d, \gamma, v_1, v_2, \alpha_1, \rho_1) = \log \gamma - \rho_1 \log \alpha_1 - (1 - \rho_1) \log(1 - \alpha_1)$$

which is at least bounded for fixed values of the hyperparameters. Also, for fixed ρ_1, γ and d, Φ is minimised: i.e. most inclined to combine when

$$v_1 + \alpha_1 d = \rho_1 \quad \text{and} \quad v_2 + \alpha_2 d = \rho_2$$

This is satisfied when $\rho_1^{-1} u_1 = \rho_2^{-1} u_2$ and $\rho_1^{-1} \alpha_1 = \rho_2^{-1} \alpha_2$. For two clusters of moderate cardinality and moderate settings of hyperparameters, this implies that we are most inclined to combine when the sample variances of the two clusters are approximately equal, and when they are at a distance from zero consistent with being from the same distribution respectively. Both these properties are clearly desirable. If we choose not to set $b_c = \breve{b} n_c$ but just to a common value for all clusters, as is done in Heard et al. (2006), then if either n_1 or n_2 is large then $v_1 + v_2 \simeq 1$ and so little distortion from a separation is felt. Two clusters with low cardinality, small values of d, and small but proportionately different values of sums of squares tend to be kept separate. For the purposes of gene clustering this is not a particularly germane property since such clusters are unlikely to be of regulatory interest and also occasionally distort the initial stages of forward selection techniques like AHC.

Finally a g-prior, ensures that Φ is increasing in the angular distance $1 - \cos\theta$ between the clusters. However, without a g-prior and the more usual dependence model, with large (n_1, n_2) it is easily checked that this is still approximately the case.

Thus, whilst in our context of gene profile clustering we would recommend the use of the proportional model, provided that we set $\rho_1 + \rho_2 = 1$ for larger cluster cardinalities, the conjugate combination algorithm will often have reasonable characteristics. Any problems that arise tend to concern the combination of clusters with small cardinalities: an inevitable consequence of using algorithms like AHC, but avoidable if more sophisticated search algorithms (Chipman et al., 2001) are employed.

7.6. Setting hyperparameters in proportional models

There are two complementary and fully Bayesian ways of setting the hyperparameters $(\breve{a}, \breve{b}, \breve{g})$. First, these parameters should be chosen so as to coincide with predictive beliefs about the individual cluster profiles we expect to see before incorporating the data. The value of $\breve{a}/\breve{b}$ is our prior expectation of the precision σ^{-2} of a typical cluster, whilst $\breve{a}$ can be calibrated to our coefficient of variation of this information $[\breve{a} n_c]^{-1}$ for a cluster c of a given cardinality n_c. The magnitude of $\breve{g}$ determines the relative strength of the prior information on each unit profile and governs the extent that the cluster posterior means shrink towards zero. Note that, in agreement with Wakefield et al. (2003), we recommend setting these prior parameters so that they calibrate to pre-posterior predictions of the variance of a particular cluster.

Second, it is important that the values $(\breve{a}, \breve{b}, \breve{g})$ calibrate hyperparameters to pre-posterior beliefs about the relative probabilities of adjacent partitions after realising certain hypothetical observations. Thus, the magnitude of parameter $\breve{g}$ solely influences the relative weight we place on two clusters having different orientations of profiles. The smaller this parameter, the more likely clusters — all of whose characteristics are the same but whose orientations are different — are kept separate. That is, the higher weight given to $\Delta^{(2)}$ relative to $\Delta^{(1)}$. To fix an appropriate value of $\breve{g}$ we suggest calibrating to two expected profiles of different orientation distances from the value of $\breve{g}$ we suggest calibrating to two expected profiles of different orientation distances from the origin and asking the scientist which two profiles are most likely to come from the same cluster.

The effects of the setting of the value of $\breve{b}$ has a strong effect on the combination rule when clusters have profiles close to zero. If it is set very small so that $d \to 0$ then two clusters with small cardinality and a ratio of the sums of squares very different from unity will be kept apart. Within the context of our running example, such gene expression profiles are not in practice interesting enough to keep separate and this phenomenon can sometimes disrupt the AHC algorithm. So, at least pragmatically, there are good reasons for keeping this parameter well away from zero. This implicitly demands that the prior expectation on the precision σ^{-2} is not big: often a plausible assumption. Interestingly, this parameter is set by default to be very small in Heard et al. (2006) which may account for a different type of instability in their algorithm that sometimes occurs early in the AHC.

The effect of the parameter $\breve{a}$ is only felt through the threshold κ''. Thus for example, using the default prior capacities suggested in Heard et al. (2006) it can be shown after a little algebra that $\breve{a}$ also controls the penalty on cluster size. When, as will be most common, this prior coefficient of variation is high (so that $\breve{a}$ is small) then the penalty tends to hold more equally sized clusters apart and to pull smaller clusters into larger ones whenever possible. However note that the setting of the prior capacities also acts solely as a penalty on (n_1, n_2) and is therefore somewhat confounded with $\breve{a}$.

8. Conclusions

Our experiences suggest that simply getting hyperparameters in the right ball park as described above can dramatically improve the characteristics of these search algorithms, see Anderson et al. (2006). Conjugate models with proportional parameter settings are not only fast but, if reasonably calibrated, behave appropriately. Even the occasional outlier can be identified and easily separated from the body of a cluster, iterating on the search algorithm if this is then necessary. The inconvenience in having to do this appears to us a small price to pay for the fast conjugate algorithm.

One useful spin-off of this analysis is that we have noted that for gene regulation, after a MAP partition has been found, the between-cluster statistic η is a useful summary. Thus clusters of genes that are potentially co-regulated can be expected to have similar profile shapes whilst the extent of the expressions, as measured by (ζ, d), is less biologically significant. Note that under Bayesian selection, provided search is extensive, all subsets of genes in a cluster will have similar associated values of η to other clusters and so this parameter not only characterises differences between clusters but also differences between collections of genes *within* clusters. This stability is important in this application since certain subsets of genes within clusters are of known biological function and therefore of more interest than others and would not be accounted for by other more ad hoc methods. Note that the

separation η between any two clusters is trivial to calculate given the previously computed statistics associated with the clusters in the MAP partition.

Finally, it is important to point out that although the problems addressed in this paper are easy to *demonstrate* using a conjugate analysis, many are not simply a *consequence* of conjugacy but actually derive from a misinterpretation of a Bayes factor as a separation measure. There is every reason to believe that other non-conjugate selection based on Bayes factors and routinely chosen prior hyperparameters will also exhibit analogous unfortunate properties. Indeed, conjugate analysis has much useful symmetry which is destroyed by incorporating different priors. The effect of introducing this lack of symmetry through the use of non-conjugate models is likely to be influential to the selection, but very difficult to characterise so that the inevitably influential hyperparameters can be set appropriately. We speculate that most current numerical analogues of the models discussed here which exhibit the same qualitative hierarchical structure will be prone not to act as if guided by a separation measure. The same care is needed to ensure genuine prior predictive beliefs are specified, otherwise the *formal* selection (and not just its numerical approximation) is likely to be unstable in this more general setting.

9. Acknowledgements

Paul E. Anderson thanks BioSim, EPSRC and BBSRC and Silvia Liverani and Jim Q. Smith thank for support from EPSRC under CRiSM. We thank Nick Heard and Chris Holmes for valuable discussions and the use of their publicly available code and also Andrew J. Millar and Kieron D. Edwards for producing the biological data shown in figures 1 and 2.

Appendix A

A.1. Proof of Theorem 1

If $\mathbf{x}(c_1) = \mathbf{x}(c_2)$ then $\alpha_1 = \alpha_2 = 0.5$, $v_1 = v_2 = v$ (say), $n_1 = n_2 = n$ and $\rho_1 = \rho_2 = \rho$

$$\exp(\Phi) = 2^{2\rho}\frac{1+\gamma d}{(2v+d)^{2\rho}}$$

where

$$\gamma = 0.5\left(\lambda_1 - 2\lambda_{12} + \lambda_2\right) = \left[(\overline{g}+2n)g_1\right]^{-1}(g+n)\overline{g}$$

Clearly, Φ is a function of z unless $\rho = 0.5$ implying $\bar{a} = 2a_1$. Substituting gives

$$\exp(\Phi) = 2\frac{1+\gamma d}{2v+d}$$

Since by definition v and d are functionally independent, we therefore must have

$$v = 0.5 \;\Leftrightarrow\; \bar{b} = 2b_1$$

and also

$$\gamma = 1 \;\Leftrightarrow\; 1 + \frac{n}{g_1} = 1 + \frac{2n}{\overline{g}} \;\Leftrightarrow\; \overline{g} = 2g_1$$

as required. Finally, under these conditions when $\mathbf{x}(c_1) = \mathbf{x}(c_2)$, $\Phi(\mathbf{x}(c_1), \mathbf{x}(c_2)) = \log 2$.

A.2. Proof of Theorem 2

$$\begin{aligned}\Phi &= \log(1+\gamma d) - \rho_1 \log(v_1 + \alpha_1 d) - \rho_2 \log(v_2 + \alpha_2 d) \\ &= \Delta^{(1)}(\gamma, d) + \Delta^{(2)}(v_1, \rho_1, \alpha_1, d) + I(\rho)\end{aligned}$$

where

$$\Delta^{(1)}(\gamma, d) = \log(1+\gamma d) - \log(1+d) \geq 0$$

since

$$\gamma = \lambda_1\alpha_1 - 2\lambda_{12}\sqrt{\alpha_1\alpha_2} + \lambda_2\alpha_2 \geq \gamma_0$$

with equality if and only if $\cos\theta[\mathbf{z}_1, \mathbf{z}_2] = 1$ where $\gamma_0 \geq 1$ is defined above, and

$$\Delta^{(2)}(v_1, \rho_1, \alpha_1, d) = \log(1+d) - \rho_1 \log(v_1 + \alpha_1 d) - \rho_2 \log(v_2 + \alpha_2 d) - I(\boldsymbol{\rho})$$

Note that

$$\Delta^{(1)} = 0 \;\Leftrightarrow\; \rho_2\alpha_1 = \rho_1\alpha_2$$
$$\cos\theta[\mathbf{z}_1, \mathbf{z}_2] = 1 \;\Leftrightarrow\; \frac{\alpha_1}{n_1} = \frac{\alpha_2}{n_2}$$

Also, for fixed ρ_1, ρ_2 with $\rho_1 + \rho_2 = 1$, $\sum_{j=1,2} \rho_j \log x_j$ is maximised when $x_j = T\rho_j$. So letting $T = 1 + d$, $v_j = \rho_j$ gives

$$\begin{aligned}\Delta^{(2)}(v_1, \rho_1, \alpha_1, d) &\geq \log(1+d) - \rho_1 \log(\rho_1(1+d)) - \rho_2 \log\left[\rho_2(1+d)\right] - I(\boldsymbol{\rho}) \\ &= 0 \\ &= \Delta^{(2)}(\rho_1, \rho_1, \alpha_1, d)\end{aligned}$$

Therefore

$$\Phi(v_1, \rho_1, \alpha_1, d) \geq I(\rho) = \Phi(\rho_1, \rho_1, \alpha_1, d)$$

A.3. Proof of Corollary 2

The first bullet is a direct consequence of the theorem on noting that $\Delta^{(1)} \geq 0$, and $\Delta^{(1)} = 0$ takes its maximum if and only if $\gamma = 1$ and $\frac{\alpha_1}{n_1} = \frac{\alpha_2}{n_2}$ so that the scaled distances of the two profiles from zero satisfy $n_1^{-1}\widehat{\boldsymbol{\beta}}_1 = n_2^{-1}\widehat{\boldsymbol{\beta}}_2$. Also, $\Delta^{(2)} \geq 0$ and $\Delta^{(2)} = 0$ if and only if $\alpha_j = \rho_j$ so that

$$v_j + \alpha_j d = \rho_j(1+d) \;\Leftrightarrow\; v_j = \rho_j \;\Leftrightarrow\; \widehat{\sigma}_1^2 = \widehat{\sigma}_2^2$$

The second bullet is immediate from the symmetry in $(\mathbf{x}(c_1), \mathbf{x}(c_2))$ of the three functions $\Phi(\mathbf{x}_1(c_1), \mathbf{x}_2(c_2))$, $I(\boldsymbol{\rho})$ and $\kappa'(\mathbf{n})$.

References

Anderson, P. E., J. Q. Smith, K. D. Edwards, and A. J. Millar (2006). Guided Conjugate Bayesian Clustering for Uncovering Circadian Genes. Technical Report 06-07, CRiSM paper, Department of Statistics, University of Warwick.

Barry, D. and J. A. Hartigan (1992). Product partitions for change point problems. *Annals of Statistics 20*, 260–279.

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

Chipman, H., E. George, and R. McCullough (1998). Bayesian CART Model Search. *J. Amer. Statist. Assoc. 93*, 935–960.

Chipman, H. and R. Tibshirani (2006). Hybrid Hierarchical Clustering with Applications to Microarray Data. *Biostatistics 7*, 268–285.

Chipman, H. A., E. George, and R. E. McCullock (2001). The Practical Implementation of Bayesian Model Selection. *Model Selection 38*, 1–50.

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

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

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

Fernandez, C., E. Ley, and M. J. F. Steel (2001). Benchmark priors for Bayesian Model Averaging. *Journal of Econometrics 100*, 381–427.

Garthwaite, P. H. and J. H. Dickey (1992). Elicitation of prior distributions for variable-selection problems in regression. *Annals of Statistics 20*(4), 1697–1719.

Gordon, A. (1999). *Classification* (2nd ed.). CRC Press, London: Chapman and Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Heard, N. A., C. C. Holmes, and D. A. Stephens (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Applictaion of Bayesian Hierarchical Clustering of Curves. *J. Amer. Statist. Assoc. 101*(473), 18–29.

McCullagh, P. and J. Yang (2006). Stochastic classification models. In *Proceedings of the International Congress of Mathematicians*, Madrid.

O'Hagan, A. and J. Forster (2004). *Bayesian Inference: Kendall's Advanced Theory of Statistics* (2nd ed.). Arnold.

O'Hagan, A. and H. Le (1994). Conflicting Information and a Class of Bivariate Heavy-tailed Distributions. In P. R. Freeman and A. F. M. Smith (Eds.), *Aspects of Uncertainty*, pp. 311–327. Wiley.

Quintana, F. A. and P. L. Ingelias (2003). Bayesian Clustering and Product Partition Models. *J. Royal Statist. Soc.: Series B 65*(2), 557–574.

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

Smith, M. and R. Kohn (1996). Non-parametric Regression using Bayesian Variable Selection. *Journal of Econometrics 75*, 317–343.

Wakefield, J., C. Zhou, and S. Self (2003). Modelling gene expression over time: curve clustering with informative prior distributions. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7*. Oxford University Press.

West, M., , and J. Harrison (1997). *Bayesian forecasting and dynamic models* (2nd ed.). New York: Springer-Verlag.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno De Finetti*, pp. 233–243. Elsevier.