Biometrika (2006), **93,** 4, *pp.* 877–893

Printed in Great Britain

Variable selection in clustering via Dirichlet process mixture models

BY SINAE KIM

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

sinae@stat.tamu.edu

MAHLET G. TADESSE

Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6021, U.S.A.

mtadesse@cceb.upenn.edu

AND MARINA VANNUCCI

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

mvannucci@stat.tamu.edu

SUMMARY

The increased collection of high-dimensional data in various fields has raised a strong interest in clustering algorithms and variable selection procedures. In this paper, we propose a model-based method that addresses the two problems simultaneously. We introduce a latent binary vector to identify discriminating variables and use Dirichlet process mixture models to define the cluster structure. We update the variable selection index using a Metropolis algorithm and obtain inference on the cluster structure via a split-merge Markov chain Monte Carlo technique. We explore the performance of the methodology on simulated data and illustrate an application with a DNA microarray study.

Some key words: Bayesian inference; Clustering; Dirichlet process mixture model; DNA microarray data analysis; Variable selection.

1. INTRODUCTION

In recent years, high-dimensional datasets have become common in various areas of application. Often, the goal of the analysis is to uncover the group structure of the observations and identify variables that best distinguish the different groups. A typical example is the analysis of DNA microarray data, where there is interest in discovering disease subtypes and isolating discriminating genes. The results could lead to a better understanding of the underlying biological processes and help develop targeted treatment strategies.

The practical utility of variable selection is well recognised and several methods have been developed for regression and classification models; see for example George & McCulloch (1993) and Sha et al. (2004). Few contributions have been made in the context

of clustering. This is a more challenging problem since there is no observed response to guide the selection. In addition, the inclusion of unnecessary variables could complicate or mask the recovery of the clusters (Tadesse et al., 2005). Liu et al. (2003) address the problem by first reducing the dimension of the data using principal component analysis and then fitting to the factors a mixture model with a fixed number of clusters. They use Markov chain Monte Carlo sampling techniques to update the sample allocations and the number of factors deemed relevant for the clustering. In practice, however, the number of clusters is not known and there is often interest in evaluating the actual variables. The principal components, which are linear combinations of all variables, do not have a straightforward interpretation. Recently, Friedman & Meulman (2004) have proposed an algorithmic approach for clustering observations on separate subsets of variables. They formulate the problem in terms of distance-based clustering with weighted variables. They use heuristic search strategies to find an optimal weighting of the variables while jointly minimising the clustering criterion. Their approach works in conjunction with hierarchical clustering, and hence does not provide inference on the number of clusters, nor does it provide a measure of uncertainty for the sample allocations. Model-based approaches have also recently been proposed. Hoff (2006) adopts a mixture of Gaussian distributions where different clusters are identified by mean shifts. The model parameters are updated using Markov chain Monte Carlo sampling techniques and Bayes factors are computed to identify discriminating variables. Both Friedman & Meulman's and Hoff's methods allow separate subsets of variables to discriminate different groups of observations. Tadesse et al. (2005) have put forward a variable selection method in which latent variables are introduced to identify discriminating variables and the clustering is formulated in terms of a finite mixture of Gaussian distributions with an unknown number of components. They used a reversible jump Markov chain Monte Carlo technique to allow for the creation and deletion of clusters. Unlike the procedures of Friedman & Meulman and Hoff, this approach assumes that the same subsets of variables discriminate across all components. However, the variable selection technique they adopt has the advantage of allowing flexible inference on both joint and marginal posterior distributions of the variables.

In this paper, we build on the model of Tadesse et al. (2005) by formulating the clustering in terms of an infinite mixture of distributions via Dirichlet process mixtures. Samples from a Dirichlet process are discrete with probability one and can therefore produce a number of ties, thereby forming clusters.

2. Model formulation

2·1. *Clustering via Dirichlet process mixture models*

A long-standing issue in all clustering procedures, including mixture models (McLachlan & Basford, 1988; Banfield & Raftery, 1993), is the problem of determining the number of clusters. This can be handled by fitting finite mixtures with an unknown number of components, such as the reversible jump Markov chain Monte Carlo algorithm (Richardson & Green, 1997; Tadesse et al., 2005) and continuous time Markov birth-death processes (Stephens, 2000a), which allow for creation and deletion of components. An alternative approach is to define mixture distributions with a countably infinite number of components. These models can be implemented by employing a Dirichlet process prior for the mixing proportions (Antoniak, 1974; Ferguson, 1983), and various Markov chain

Monte Carlo sampling methods for fitting Dirichlet process mixture models have been developed (Escobar, 1994; MacEachern, 1994; Escobar & West, 1995; MacEachern & Müller, 1998).

Let $X = (x_1, \ldots, x_n)$ be independent p-dimensional observations arising from a mixture of distributions $F(\theta_i)$. The model parameters specific to individual i, θ_i, are assumed to be independent draws from some distribution, G, which in turn follows a Dirichlet process prior. This leads to the following hierarchical mixture model:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i), \\ \theta_i|G &\sim G, \\ G &\sim \mathrm{DP}(G_0, \alpha), \end{aligned} \tag{1}$$

where G_0 defines a baseline distribution for the Dirichlet process prior, such that $E(G) = G_0$, and α is a concentration parameter. The Pólya urn scheme representation of the Dirichlet process provides the basis for most computational strategies to fit this model (Blackwell & MacQueen, 1973). Integrating over G allows the θ_i to be written in terms of successive conditional distributions:

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha} \sum_{k \neq i} \delta(\theta_k) + \frac{\alpha}{n-1+\alpha} G_0, \tag{2}$$

where $\delta(\theta_k)$ is a point mass distribution at θ_k.

Equivalent models can be obtained by taking the limit as $K \to \infty$ of finite mixture models with K components. This leads to

$$\begin{aligned} x_i|c_i, \phi &\sim F(\phi_{c_i}), \\ c_i|p &\sim \mathrm{Discrete}(p_1, \ldots, p_K), \\ \phi_c &\sim G_0, \\ p &\sim \mathrm{Dir}(\alpha/K, \ldots, \alpha/K), \end{aligned} \tag{3}$$

where the latent variable c_i indicates the cluster allocation of sample i and ϕ_{c_i} corresponds to the identical θ_i's. As shown in Neal (2000), integrating over the mixing proportions p and taking $K \to \infty$ leads to the following prior for c_i:

$$\begin{aligned} \mathrm{pr}(c_i = c_l \text{ for some } l \neq i|c_{-i}) &= \frac{n_{-i,c_l}}{n-1+\alpha}, \\ \mathrm{pr}(c_i \neq c_l \text{ for all } l \neq i|c_{-i}) &= \frac{\alpha}{n-1+\alpha}, \end{aligned} \tag{4}$$

where $n_{-i,c}$ is the number of $c_l = c$ for $l \neq i$. Thus, sample i is allocated to an existing cluster with probability proportional to the cluster size and it is assigned to a new cluster with probability proportional to α. As shown in Antoniak (1974), the prior probability of observing exactly k distinct clusters is given by

$$\mathrm{pr}(K = k|\alpha, n) = {}_n a_k \alpha^k \frac{1}{A_n(\alpha)}, \tag{5}$$

where the coefficients ${}_n a_k$ are the absolute values of Stirling numbers of the first kind (Abramowitz & Stegun, 1964, p. 833) and $A_n(x) = {}_n a_1 x + {}_n a_2 x^2 + \ldots + {}_n a_n x^n$.

If G_0 in (3) is a conjugate prior for F, sampling from the posterior distribution using Gibbs sampling is straightforward. We will consider a procedure in which conjugacy is fully exploited as described by Neal (1992). Integrating out the model parameters ϕ_{c_i} simplifies the algorithm considerably, as the latent indicators c_i will then be the only parameters to be updated. The conditional probabilities for the c_i's are then given by

$$\begin{aligned} \text{pr}(c_i = c_l \text{ for some } l \neq i | c_{-i}, x_i) &= b \frac{n_{-i,c_l}}{n-1+\alpha} \int F(x_i; \phi) dH_{-i,c_l}(\phi), \\ \text{pr}(c_i \neq c_l \text{ for some } l \neq i | c_{-i}, x_i) &= b \frac{\alpha}{n-1+\alpha} \int F(x_i; \phi) dG_0(\phi), \end{aligned} \tag{6}$$

where b is the appropriate normalising constant, and $H_{-i,c}$ is the posterior distribution of ϕ based on the prior G_0 and all observations x_l for which $l \neq i$ and $c_l = c$.

The Gibbs sampler and the sequential importance sampling (MacEachern et al., 1999), which rely on the Pólya-urn-based incremental update, suffer from slow mixing. Several methods have been developed to overcome this problem. One such approach is the blocked Gibbs sampler of Ishwaran & James (2001) which updates blocks of parameters. Green & Richardson (2001) have proposed the use of split/merge moves in the spirit of their reversible jump procedure for finite mixture models (Richardson & Green, 1997). Jain & Neal (2004) and Dahl (2006) have also proposed sampling schemes that involve splitting and merging of clusters to circumvent the lack of mixing of the standard Gibbs sampler. Here, we make use of Jain & Neal's (2004) split-merge Markov chain Monte Carlo procedure. The method, which is described in § 2·3, escapes local modes by separating or combining a group of observations based on the Metropolis–Hastings algorithm.

2·2. *Variable selection in clustering*

Unlike linear models and classification problems, where the response variable is observed and guides the selection, here the sample allocations are unknown parameters that need to be estimated. Stochastic search variable selection techniques (George & McCulloch, 1993; Brown et al., 1998) have been used successfully in various applications to identify informative predictors. These methods introduce a latent binary vector γ to index all possible models and use the γ_j's to induce a mixture prior on the corresponding regression coefficients. However, clustering is different from a regression setting and the following adjustment is needed to define the latent indicators (Tadesse et al., 2005):

$$\gamma_j = \begin{cases} 1, & \text{if variable } j \text{ defines a mixture distribution,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The latent vector γ is therefore used to identify directly variables that discriminate between the different groups. We denote by $X_{(\gamma)}$ the set of variables that define mixture distributions and by $X_{(\gamma^c)}$ the remaining variables which favour one multivariate density across all observations.

Our goal is to combine the clustering and variable selection tasks. We assume that $F(\phi_{c_i})$ in (3) is an infinite mixture of Gaussian distributions with component parameters $\phi_k = (\mu_k, \Sigma_k)$. Thus, conditional on the discriminating variables, we have

$$x_{i(\gamma)} | c_i = k, \phi_k, \gamma \sim \mathcal{N}(\mu_{k(\gamma)}, \Sigma_{k(\gamma)}) \tag{8}$$

and, with $\psi = (\eta, \Omega)$, the nondiscriminating variables follow

$$x_{i(\gamma^c)}|\psi, \gamma \sim \mathcal{N}(\eta_{(\gamma^c)}, \Omega_{(\gamma^c)}). \tag{9}$$

The likelihood function therefore consists of the contribution from the clustering and nonclustering covariates which, assuming no correlation between the two sets of variables, is given by

$$\begin{aligned}\mathcal{L}(c, \gamma, \phi, \psi|X) = (2\pi)^{-\{n(p-p_\gamma)\}/2}|\Omega_{(\gamma^c)}|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_{i(\gamma^c)} - \eta_{(\gamma^c)})^{\mathrm{T}}\Omega_{(\gamma^c)}^{-1}(x_{i(\gamma^c)} - \eta_{(\gamma^c)})\right\} \\ \times \prod_{k=1}^{K}(2\pi)^{-(n_k p_\gamma)/2}|\Sigma_{k(\gamma)}|^{-n_k/2} \exp\left\{-\frac{1}{2}\sum_{i \in C_k}(x_{i(\gamma)} - \mu_{k(\gamma)})^{\mathrm{T}}\Sigma_{k(\gamma)}^{-1}(x_{i(\gamma)} - \mu_{k(\gamma)})\right\},\end{aligned} \tag{10}$$

where $p_\gamma = \sum_{j=1}^{p}\gamma_j$ and $C_k = \{i : c_i = k, i = 1, \ldots, n\}$ with cardinality n_k. In practice, it is plausible to have correlation between clustering and nonclustering variables, but it is difficult to accommodate such structure. Our assumption provides computational convenience. It is well known that high correlation among covariates complicates data analysis. In our approach, since the component parameters are integrated out and not estimated, we believe that the implications of ignoring the correlation between the two sets of variables will be minimal.

For the prior specification on γ, we consider its elements, γ_j, to be independent Bernoulli random variables with common probability

$$\mathrm{pr}(\gamma) = \prod_{j=1}^{p}\omega^{\gamma_j}(1-\omega)^{1-\gamma_j}, \tag{11}$$

where ω can be elicited as the proportion of variables expected a priori in the discriminating set. Any further knowledge about some of the variables or their interactions can be incorporated in the prior.

As mentioned above we specify conjugate priors and integrate out the mean and covariance parameters. For computational convenience, we assume independence among the nondiscriminating variables and set $\Omega = \sigma^2 I_{p\times p}$. We specify the prior distributions as follows:

$$\begin{gathered}\mu_{k(\gamma)}|\Sigma_{k(\gamma)} \sim \mathcal{N}(\mu_{0(\gamma)}, h_1\Sigma_{k(\gamma)}), \quad \eta_{(\gamma^c)}|\Omega_{(\gamma^c)} \sim \mathcal{N}(\mu_{0(\gamma^c)}, h_0\Omega_{(\gamma^c)}), \\ \Sigma_{k(\gamma)} \sim \mathrm{IW}(\delta; Q_{1(\gamma)}), \quad \sigma^2 \sim \mathrm{IG}(a, b),\end{gathered} \tag{12}$$

where $\mathrm{IW}(\delta; Q_1)$ is an inverse Wishart distribution with dimension p, shape parameter $\delta = n - p + 1$, n degrees of freedom and mean $Q_1/(\delta - 2)$ (Brown, 1993, Appendix A). The notation $\mathrm{IG}(a, b)$ denotes an inverse gamma distribution with mean $b/(a-1)$ and variance $b^2/\{(a-1)^2(a-2)\}$. Small values of δ lead to weak prior information. We set $\delta = a = 3$, the smallest integer such that the mean and variance of the corresponding densities are defined, and take $Q_1 = \kappa_1 I_{p\times p}$. Some care is needed in the choice of κ_1 and b. These hyperparameters need to be specified in the range of variability of the data. We found values close to the mean variance of the columns of X to yield reasonable results. For the mean parameters, we take the priors to be fairly flat over the region where the data are defined. Each element of μ_0 is set to the corresponding covariate interval midpoint. Values of h_0 and h_1 between 10 and 1000 performed well. These data-based priors ensure that the prior distributions overlap with the likelihood and that we obtain well-behaved

posterior densities. As mentioned in Richardson & Green (1997), in mixture models it is not possible to be fully noninformative and obtain proper posterior distributions. This point is also emphasised by Wasserman (2000), who proposed data-dependent priors in the context of finite mixtures. A comprehensive discussion about various prior specifications and their effects is provided in Kass & Wasserman (1996). The authors argue that the use of diffuse proper priors in complex statistical models can lead to posteriors with undesirable properties.

After the component parameters are integrated out, the marginalised likelihood becomes

$$f(X|\gamma, c) = \pi^{-np/2} \prod_{k=1}^{K} \{H_{k(\gamma)} |Q_{1(\gamma)}|^{(\delta + p_\gamma - 1)/2} |Q_{1(\gamma)} + S_{k(\gamma)}|^{-(n_k + \delta + p_\gamma - 1)/2}\} H_{0(\gamma^c)} (S_{0(\gamma^c)})^{-(a + n/2)}, \tag{13}$$

in which

$$H_{k(\gamma)} = (h_1 n_k + 1)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\{(n_k + \delta + p_\gamma - j)/2\}}{\Gamma\{(\delta + p_\gamma - j)/2\}},$$

$$H_{0(\gamma^c)} = (h_0 n + 1)^{-(p - p_\gamma)/2} b^{a(p - p_\gamma)} \prod_{j=1}^{p - p_\gamma} \frac{\Gamma(a + n/2)}{\Gamma(a)},$$

$$S_{k(\gamma)} = \sum_{i \in C_k} (x_{i(\gamma)} - \bar{x}_{k(\gamma)})(x_{i(\gamma)} - \bar{x}_{(k(\gamma)})^{\mathrm{T}} + \frac{n_k}{h_1 n_k + 1} (\mu_{0(\gamma)} - \bar{x}_{k(\gamma)})(\mu_{0(\gamma)} - \bar{x}_{k(\gamma)})^{\mathrm{T}},$$

$$S_{0(\gamma^c)} = \prod_{j=1}^{p - p_\gamma} \left[b + \frac{1}{2} \left\{ \sum_{i=1}^{n} (x_{ij(\gamma^c)} - \bar{x}_{j(\gamma^c)})^2 + \frac{n}{h_0 n + 1} (\mu_{0j(\gamma^c)} - \bar{x}_{j(\gamma^c)})^2 \right\} \right],$$

where $\bar{x}_{k(\gamma)}$ is the sample mean of cluster k, and $\bar{x}_{j(\gamma^c)}$ is the sample mean of the jth nondiscriminating variable.

2·3. *Model fitting*

We update the variable selection index using repeated Metropolis steps and carry out inference on the cluster structure using the Jain & Neal (2004) split-merge algorithm. Our procedure iterates between the following steps.

Step 1. Update the latent variable selection indicator γ by repeating the following Metropolis step t times. A new candidate γ^{new} is generated by randomly choosing one of two transition moves:

(i) add/delete by randomly picking one of the p indices in γ^{old} and changing its value;
(ii) swap by drawing independently and at random a 0 and a 1 in γ^{old} and switching their values.

The new candidate is accepted with probability

$$\min \left\{ 1, \frac{f(\gamma^{\mathrm{new}}|X, c)}{f(\gamma^{\mathrm{old}}|X, c)} \right\}, \tag{14}$$

where $f(\gamma|X, c) \propto f(X|\gamma, c)\, \mathrm{pr}(\gamma)$. This stochastic update was suggested for model selection by Madigan & York (1995) and has been used extensively for variable selection in linear models by George & McCulloch (1997), among others, and in classification by Sha et al. (2004). In the context of clustering, we are dealing with a more complex model where there is no observed outcome to guide the selection. Instead, the variable selection and

the cluster structure evolve simultaneously. Therefore, to allow the selection to stabilise for a given cluster configuration, we repeat the Metropolis steps a number of times. In general, we found little improvement in the Markov chain Monte Carlo performance beyond 20 intermediate Metropolis steps.

Step 2. Update the latent sample allocation vector c using Jain & Neal's (2004) split-merge procedure, as follows. Start by selecting two distinct observations, i and l at random uniformly. Let $\mathscr{C}$ denote the set of observations, $k \in \{1, \ldots, n\}$, for which $k \neq i$, $k \neq l$ and $c_k = c_i$ or $c_k = c_l$.

Case 1. If $\mathscr{C}$ is empty, use the following simple random split-merge algorithm.

(a) If $c_i = c_l$, then

(i) the component is split such that a new component $c_i^{\text{split}} \notin \{c_1, \ldots, c_n\}$ is created, the allocations for the other observations remaining unchanged;

(ii) the proposal is accepted with probability

$$a(c^{\text{split}}, c) = \min\left\{1, \frac{q(c|c^{\text{split}})\,\text{pr}(c^{\text{split}})L(c^{\text{split}}|X, \gamma)}{q(c^{\text{split}}|c)\,\text{pr}(c)L(c|X, \gamma)}\right\},$$

where

$$\frac{q(c|c^{\text{split}})}{q(c^{\text{split}}|c)} = 1, \quad \frac{\text{pr}(c^{\text{split}})}{\text{pr}(c)} = \alpha,$$

$$\begin{aligned}\frac{L(c^{\text{split}}|X, \gamma)}{L(c|X, \gamma)} &= \frac{\{\int F(x_i; \phi, \gamma)dG_0(\phi, \gamma)\}\{\int F(x_l; \phi, \gamma)dG_0(\phi, \gamma)\}}{\int F(x_i; \phi, \gamma)F(x_l; \phi, \gamma)dG_0(\phi, \gamma)} \\ &= \frac{(1+2h_1)^{p_\gamma/2}}{(1+h_1)^{p_\gamma}} \times \frac{|Q_{1(\gamma)}|^{(\delta+p_\gamma-1)/2}|Q_{1(\gamma)} + S_{il(\gamma)}|^{(\delta+p_\gamma+1)/2}}{(|Q_{1(\gamma)} + S_{i(\gamma)}||Q_{1(\gamma)} + S_{l(\gamma)}|)^{(\delta+p_\gamma)/2}} \\ &\quad \times \prod_{j=1}^{p_\gamma} \frac{[\Gamma\{(\delta+p_\gamma+1-j)/2\}]^2}{\Gamma\{(\delta+p_\gamma-j)/2\}\Gamma\{(\delta+p_\gamma+2-j)/2\}}, \end{aligned} \tag{15}$$

$$S_{i(\gamma)} = (1+h_1)^{-1}(x_{i(\gamma)} - \mu_{0(\gamma)})(x_{i(\gamma)} - \mu_{0(\gamma)})^{\text{T}},$$

$$\begin{aligned}S_{il}(\gamma) = (1+2h_1)^{-1}\{&(x_{i(\gamma)} - \mu_{0(\gamma)})(x_{i(\gamma)} - \mu_{0(\gamma)})^{\text{T}} + (x_{l(\gamma)} - \mu_{0(\gamma)})(x_{l(\gamma)} - \mu_{0(\gamma)})^{\text{T}} \\ &+ h_1(x_{i(\gamma)} - x_{l(\gamma)})(x_{i(\gamma)} - x_{l(\gamma)})^{\text{T}}\}.\end{aligned}$$

(b) If $c_i \neq c_l$, then

(i) c_i and c_l are merged into a single component, c^{merge};

(ii) the proposal is accepted with probability

$$a(c^{\text{merge}}, c) = \min\left\{1, \frac{q(c|c^{\text{merge}})\,\text{pr}(c^{\text{merge}})L(c^{\text{merge}}|X, \gamma)}{q(c^{\text{merge}}|c)\,\text{pr}(c)L(c|X, \gamma)}\right\},$$

where

$$\frac{q(c|c^{\text{merge}})}{q(c^{\text{mrge}}|c)} = 1, \quad \frac{\text{pr}(c^{\text{merge}})}{\text{pr}(c)} = \frac{1}{\alpha},$$

$$\frac{L(c^{\text{merge}}|X, \gamma)}{L(c|X, \gamma)} = \frac{\int F(x_i; \phi, \gamma)F(x_l; \phi, \gamma)dG_0(\phi, \gamma)}{\{\int F(x_i; \phi, \gamma)dG_0(\phi, \gamma)\}\{\int F(x_l; \phi, \gamma)dG_0(\phi, \gamma)\}}.$$

Case 2. If $\mathscr{C}$ is not empty, the following restricted Gibbs sampling split-merge is used

(a) Start by building a launch state as follows:
  (i) if $c_i = c_l$, then split the component such that $c_i^{\text{launch}} \notin \{c_1, \ldots, c_n\}$ and $c_l^{\text{launch}} = c_l$;
  (ii) if $c_i \neq c_l$, then $c_i^{\text{launch}} = c_i$ and $c_i^{\text{launch}} = c_l$;
  (iii) for every $k \in \mathscr{C}$, set c_k^{launch} independently and at random with probability 0·5 to either c_i^{launch} or c_l^{launch};
  (iv) perform t intermediate restricted Gibbs sampling scans to allocate each observation $k \in \mathscr{C}$ to either c_i^{launch} or c_l^{launch} using the conditional distribution

$$\text{pr}(c_k | c_{-k}, x_k, \gamma) = \frac{n_{-k,c_k} \int F(x_k; \phi, \gamma) dH_{-k,c_k}(\phi, \gamma)}{n_{-k,c_i} \int F(x_k; \phi, \gamma) dH_{-k,c_i}(\phi, \gamma) + n_{-k,c_l} \int F(x_k; \phi, \gamma) dH_{-k,c_l}(\phi, \gamma)}, \quad (16)$$

where

$$\int F(x_k; \phi, \gamma) dH_{-k,c_i}(\phi, \gamma) = \pi^{-p_\gamma/2} \left(\frac{h_1 n_{c_i} + 1}{h_1 n_{-k,c_i} + 1} \right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\{(n_{c_i} + \delta + p_\gamma - j)/2\}}{\Gamma\{(n_{-k,c_i} + \delta + p_\gamma - j)/2\}} \times |Q_{1(\gamma)} + S_{c_i(\gamma)}|^{-(n_{c_i} + \delta + p_\gamma - 1)/2} |Q_{1(\gamma)} + S_{-k,c_i(\gamma)}|^{(n_{-k,c_i} + \delta + p_\gamma - 1)/2},$$

with

$$S_{-k,c_i(\gamma)} = \sum_{j \neq k : c_j = c_i} (x_{j(\gamma)} - \bar{x}_{c_i(\gamma)})(x_{j(\gamma)} - \bar{x}_{c_i(\gamma)})^{\mathrm{T}} + \frac{n_{-k,c_i}}{h_1 n_{-k,c_i} + 1} (\mu_{0(\gamma)} - \bar{x}_{c_i}(\gamma))(\mu_{0(\gamma)} - \bar{x}_{c_i(\gamma)})^{\mathrm{T}}$$

and $S_{c_i(\gamma)}$ is defined as in equation (13).

Jain & Neal (2004) found that the improvement in mixing is minimal after five intermediate scans. The result from the last restricted Gibbs sampling scan constitutes the launch state for the split-merge procedure.

(b) If $c_i = c_l$, then
  (i) let $c_i^{\text{split}} = c_i^{\text{launch}}$ and $c_l^{\text{split}} = c_l^{\text{launch}}$;
  (ii) for every observation $k \in \mathscr{C}$, perform one final Gibbs sampling scan from c^{launch} to set c_k^{split} to either c_i^{split} or c_l^{split} using equation (16);
  (iii) the allocation for observations $k \notin \mathscr{C} \cup \{i, l\}$ remains unchanged, $c_k^{\text{split}} = c_k$;
  (iv) evaluate the proposal by the Metropolis–Hastings acceptance probability $a(c^{\text{split}}, c)$, where $q(c^{\text{split}}|c)$ is obtained by computing the Gibbs sampling transition probability from c^{launch} to c^{split}.

(c) If $c_i \neq c_l$, then
  (i) let $c_i^{\text{merge}} = c_l$ and $c_l^{\text{merge}} = c_l$;
  (ii) for every observation $k \in \mathscr{C}$, let $c_k^{\text{merge}} = c_l$;
  (iii) the allocation for observations $k \notin \mathscr{C} \cup \{i, l\}$ remains unchanged, $c_k^{\text{merge}} = c_k$;
  (iv) the proposal is accepted with probability $a(c^{\text{merge}}, c)$, where $q(c|c^{\text{merge}})$ is the product over $k \in \mathscr{C}$ of the probabilities of setting each c_k in the original split state to its value in the launch state.

One iteration is completed after performing a full Gibbs sampling scan and updating all sample allocations c_i $(i = 1, \ldots, n)$ from their conditional distributions given by

$$\begin{aligned}\mathrm{pr}(c_i = c_l \text{ for some } l \neq i | c_{-i}, X, \gamma) &\propto \pi^{-np_\gamma/2} \frac{n_{-i,c_l}}{n-1+\alpha} \left(\frac{h_1 n_{c_l} + 1}{h_1 n_{-i,c_l} + 1} \right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\{(n_{c_l} + \delta + p_\gamma - j)/2\}}{\Gamma\{(n_{-i,c_l} + \delta + p_\gamma - j)/2\}} \\ &\quad \times |Q_{1(\gamma)} + S_{c_l(\gamma)}|^{-(n_{c_l} + \delta + p_\gamma - 1)/2} |Q_{1(\gamma)} + S_{-i,c_l(\gamma)}|^{(n_{-i,c_l} + \delta + p_\gamma - 1)/2}, \end{aligned} \tag{17}$$

where

$$S_{k(\gamma)} = \sum_{l:c_l = k} (x_{l(\gamma)} - \bar{x}_{k(\gamma)})(x_{l(\gamma)} - \bar{x}_{k(\gamma)})^{\mathrm{T}} + \frac{n_k}{h_1 n_k + 1} (\mu_{0(\gamma)} - \bar{x}_{k(\gamma)})(\mu_{0(\gamma)} - \bar{x}_{k(\gamma)})^{\mathrm{T}}$$

and $\mathrm{S}_{-i,k(\gamma)}$ is the equivalent expression without the ith observation, and

$$\begin{aligned}\mathrm{pr}(c_i \neq c_l \text{ for all } l \neq i | c_{-i}, X, \gamma) &\propto \pi^{-np_\gamma/2} \frac{\alpha}{n-1+\alpha} (h_1 + 1)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\{(1 + \delta + p_\gamma - j)/2\}}{\Gamma\{(\delta + p_\gamma - j)/2\}} \\ &\quad \times |Q_{1(\gamma)}|^{(\delta + p_\gamma - 1)/2} |Q_{1(\gamma) + S_{i(\gamma)}}|^{-(\delta + p_\gamma)/2}, \end{aligned} \tag{18}$$

where $S_{i(\gamma)} = (h_1 + 1)^{-1}(x_{i(\gamma)} - \mu_{0(\gamma)})(x_{i(\gamma)} - \mu_{0(\gamma)})^{\mathrm{T}}$.

The split-merge algorithm helps improve the mixing of the sampler, which is a typical problem in fitting mixture models. The problem here is further aggravated by the inclusion of variable selection. In cases where the sampler still exhibits poor performance, becoming stuck at a local mode and not accepting the proposed split-merge moves, a tempering scheme can be introduced. One such approach is the parallel tempering algorithm (Geyer, 1991). A series of distributions that interpolate between the distribution of interest and a distribution from which sampling is easier are defined, such that $f_t(c|X, \gamma) = f(c|X, \gamma)^{1/T_t}$, for $t = 1, \ldots, T$. The procedure consists of the following steps.

Step 1: Parallel scan. For each chain with equilibrium distribution $f_t(.)$, $c^{\mathrm{old}}(T_t)$ is updated to $c^{\mathrm{new}}(T_t)$ as described above.

Step 2: State exchange. Two neighbouring chains, T_t and $T_{t'}$, are randomly chosen and an attempt is made to swap $c^{\mathrm{new}}(T_t)$ with $c^{\mathrm{new}}(T_{t'})$. This update is accepted with probability

$$\min \left\{ 1, \left[\frac{f\{c^{\mathrm{new}}(T_{t'})|X, \gamma\}}{f\{c^{\mathrm{new}}(T_t)|X, \gamma\}} \right]^{(T_t^{-1} - T_{t'}^{-1})} \right\}.$$

3. Posterior inference

3·1. *Inference about c*

For inference about the cluster structure, a commonly used estimate is the maximum a posteriori sample allocation vector, which corresponds to the configuration with highest conditional posterior probability among those drawn by the Markov chain Monte Carlo sampler:

$$\hat{c} = \underset{1 \leqslant t \leqslant M}{\mathrm{argmax}}\ \mathrm{pr}(c^{(t)}|X, \hat{\gamma}), \tag{19}$$

where $\hat{\gamma}$ is the set of variables selected based on the marginal posterior probabilities $\text{pr}(\gamma_j = 1|X)$.

We also investigate another estimator that relies on the posterior pairwise probabilities, $\text{pr}(c_i = c_j|X)$, estimated by the empirical frequencies in the Markov chain Monte Carlo output. With a sample size n there are $\binom{n}{2}$ such pairwise posterior probabilities, which can be viewed as entries of a symmetric $n \times n$ similarity matrix. An approach proposed by Dahl (2006), which he refers to as least-squares clustering, estimates the cluster structure by forming an association matrix at every Markov chain Monte Carlo iteration. Each cell of the association matrix takes the value 1 if the corresponding row and column elements are allocated to the same cluster and 0 otherwise. The sum of absolute deviations between the entries of the association matrix and those of the similarity matrix is then calculated for each Markov chain Monte Carlo output, and the configuration that minimises that quantity is considered.

3·2. *Inference about* γ

Inference about variables that discriminate between the different groups can be done either through the joint posterior distribution of γ or through the marginal posterior distributions of its elements. The former selects variables based on

$$\hat{\gamma} = \underset{1 \leqslant t \leqslant M}{\text{argmax}}\ \text{pr}(\gamma^{(t)}|X, \hat{c}), \tag{20}$$

where $\hat{c}$ is the sample allocation estimated based on $\text{pr}(c_i = c_j|X)$. The latter identifies the variables with largest marginal posterior probabilities $\text{pr}(\gamma_j = 1|X)$, which are estimated by the empirical frequencies in the Markov chain Monte Carlo output.

4. Data analysis

4·1. *Simulation study*

We first investigate the performance of the methodology using simulated data. We generate a dataset of 15 observations and 1000 variables, where a set of 20 variables are chosen to separate the observations into four components:

$$\begin{aligned} x_{ij} \sim\ & I_{\{1 \leqslant i \leqslant 4\}} N(\mu_1, \sigma_1^2) + I_{\{5 \leqslant i \leqslant 7\}} N(\mu_2, \sigma_2^2) + I_{\{8 \leqslant i \leqslant 13\}} N(\mu_3, \sigma_3^2) \\ & + I_{\{14 \leqslant i \leqslant 15\}} N(\mu_4, \sigma_4^2) \quad (i = 1, \ldots, 15,\ j = 1, \ldots, 20), \end{aligned} \tag{21}$$

where $I_{\{.\}}$ is the indicator function, equal to 1 if the condition is satisfied. Thus, the first four observations are generated from one group, the next three come from the second group, the next six are in the third group, and the last two fall in the fourth group. The component parameters μ_k and σ_k^2, for $k = 1, \ldots, 4$, are randomly chosen from $[-5, 5]$ and $[0{\cdot}01, 1]$ respectively. The remaining 980 variables, which do not separate the samples into clusters, are drawn from a standard normal density.

We chose the hyperparameters h_1 and κ_1 such that $h_1 \times \kappa_1$ is close to the mean of the empirical variances from the p variables. We set $h_1 = 1000$ and found the results to be quite robust for values of κ_1 in the range $[5 \times 10^{-4}, 2 \times 10^{-3}]$. For the nondiscriminating variables, we chose b equal to the mean of the variances and found h_0 values between 10 and 100 to perform well. We report here the results for $\alpha = 1$, $\delta = a = 3$, $\kappa_1 = 7 \times 10^{-4}$,

$h_0 = 100$, $b = 0{\cdot}2$ and $\omega = 10/p$. We started a Markov chain Monte Carlo run from a vector γ with 10 randomly selected elements set to 1 and each observation in a separate cluster. We ran 100 000 iterations and used the first 40 000 as burn-in. At each iteration, we performed 20 repeated Metropolis steps to update γ and three restricted Gibbs scans with one final Gibbs sampling to update c. We also used the parallel tempering algorithm with two temperature ladders to improve further the mixing of the sampler. The temperatures were chosen such that the acceptance rates for exchanges between neighbouring chains are between 0·5 and 0·7.

Figures 1(a) and (b) show respectively the trace plots for the number of clusters and the number of discriminating variables. The sampler stabilised quickly around models with 3 to 5 clusters and 15 to 20 discriminating variables. We estimated the cluster allocations as described in § 3. The posterior sample allocations estimated using equation (19) favoured five components with the last two observations assigned to separate clusters. The allocations obtained using the pairwise probability estimates and the sum of absolute deviations algorithm perfectly matched the true cluster structure. Figure 2(a) displays the pairwise posterior probabilities, $\mathrm{pr}(c_i = c_j|X)$, of allocating observations i and j to the same cluster. The groupings used to simulate the data are successfully identified. For the variable selection, we ordered the visited vectors $\gamma^{(t)}$ according to their posterior probabilities and identified the 'best' subset as the $\hat{\gamma}$ from equation (20). This vector contained 17 variables, all of which are among the 20 discriminating covariates used to simulate the data. We also looked at the marginal posterior probabilities, $\mathrm{pr}(\gamma_j = 1|X)$, which are displayed in Fig. 2(b). The x-axis in this plot corresponds to the variable indices and the spikes indicate variables that have high posterior probabilities. The same 17 variables were selected at a marginal probability threshold of 0·7.

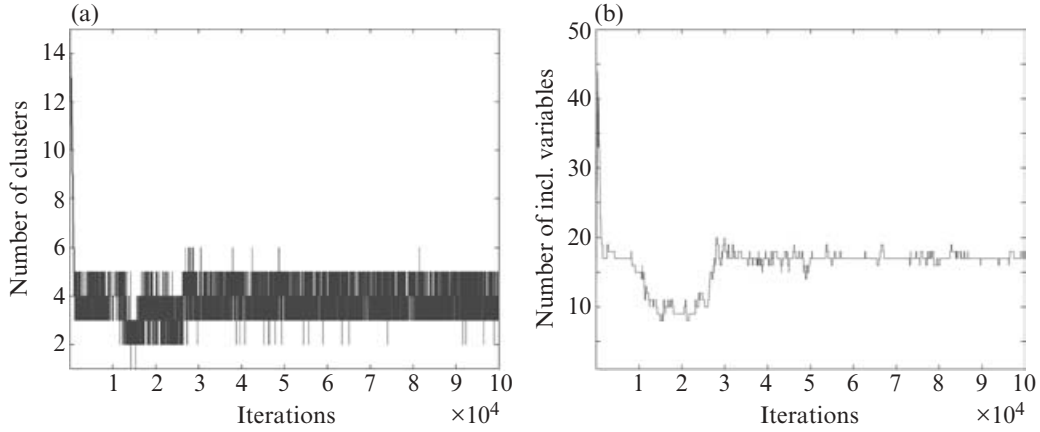

Fig. 1: Simulated data with $\alpha = 1$ and $\omega = 10/p$. Trace plots for (a) number of clusters, (b) number of included discriminating variables.

We investigated the sensitivity of the results to the choice of α and ω, which respectively influence the number of clusters and the number of selected variables. In general, we found the results to be quite robust to the values of these hyperparameters. Here, we report the results for two different choices of each parameter. We took $\alpha = 1$ and $\alpha = 15$, the latter of which is equal to the sample size. As shown in equation (5), the number of clusters is defined a priori by the sample size n in the data and the choice of the hyperparameter α. With $\alpha = 1$, the prior predictive distribution of the number of components turns out to

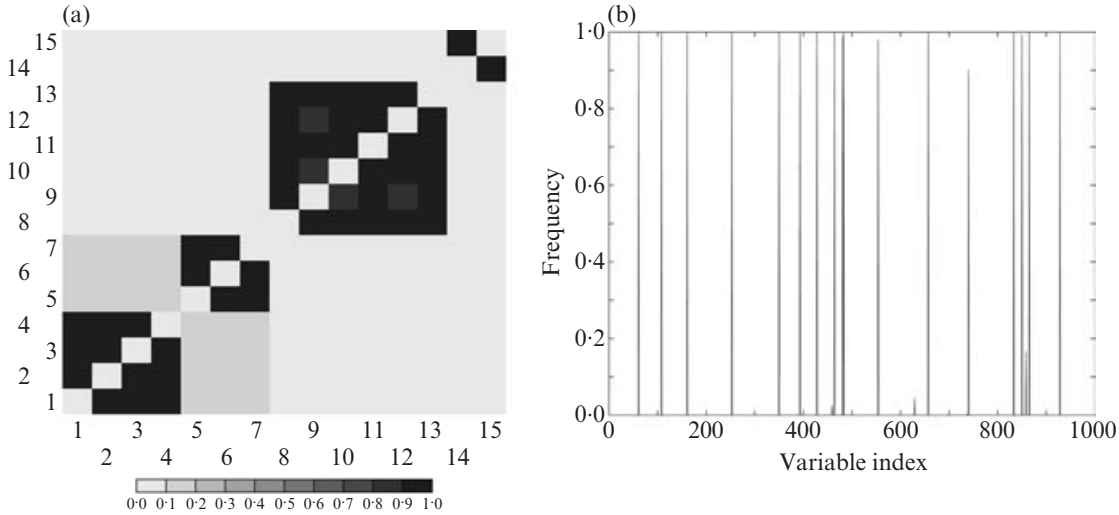

Fig. 2: Simulated data with $\alpha = 1$ and $\omega = 10/p$. (a) Pairwise posterior probabilities $\text{pr}(c_i = c_j|X)$, (b) marginal posterior probabilities $\text{pr}(\gamma_j = 1|X)$.

be concentrated between one and six, whereas with $\alpha = 15$ between 7 and 14 clusters are expected a priori. For the variable selection hyperparameter, we chose $\omega = 10/p$ and $\omega = 30/p$. The trace plots for the corresponding Markov chain Monte Carlo output, not shown, indicate that the inference on both the cluster structure and the selected variables is similar to that for the case $\alpha = 1$ and $\omega = 10/p$. However, a large value of α does make the sampler visit models with more components, although there is still strong support for models with three to five clusters. The four clusters are successfully identified and the same 15 discriminating variables are selected. A larger value of α also affects the mixing of the sampler in terms of the variable selection; this is not surprising since the cluster structure and the variable selection evolve simultaneously.

This simulated dataset is identical to the one used in Tadesse et al. (2005), where a finite mixture model with the reversible jump Markov chain Monte Carlo technique was used to infer the cluster structures, and performed much better than Friedman & Meulman's (2004) clustering objects on subsets of attributes algorithm, which implements variable selection in the context of hierarchical clustering.

4·2. *DNA microarray data analysis*

A typical application where clustering has become a common task is the analysis of DNA microarray data, where thousands of gene expression levels are monitored on a few experimental units. For example, Medvedovic & Sivaganesan (2002) used Dirichlet process mixture models to cluster genes with similar expression patterns. Our goal here is different. We want to uncover subclasses among the experimental units and identify genes that best discriminate between the different groups. This could help identify disease subtypes and understand some of the heterogeneity in treatment outcome for patients receiving similar diagnoses.

We illustrate our methodology using the widely analysed leukaemia data of Golub et al. (1999) and focus on the 38 patients from the training set. We followed the same pre-processing as other investigators (Dudoit et al., 2002) by truncating expression measures

beyond the threshold of reliable detection at 100 and 16 000, and by removing probe sets with intensities such that $\max/\min \leqslant 5$ and $\max - \min \leqslant 500$. This left 3571 genes for analysis. The expression readings were log-transformed and each variable was rescaled by its range.

We chose the hyperparameters using similar guidelines to those of the simulated example. We performed Markov chain Monte Carlo runs with α set to 1 and 38. The other hyperparameters were taken to be $\delta = a = 3$, $h_0 = 100$, $h_1 = 10$, $\kappa_1 = 0{\cdot}06$, $b = 0{\cdot}1$ and $\omega = 20/p$. For both values of α, we ran two chains with different initial models, one in which all γ_j's except one are set to 0, and the other in which 10 randomly chosen γ_j's are set to 1. In all cases, the sampler was started with all observations assigned to one cluster and 200 000 iterations were run with the first 100 000 used as burn-in.

Figures 3(a) and (b) give the summary trace plots for the number of clusters and the number of discriminating variables using $\alpha = 38$ for one of the chains. The sampler mixed well mostly visiting models with four to seven components. As for the number of variables, the chain stabilised near models with 120 discriminating variables. The second run gave similar results. For posterior inference, we pooled the output from the two chains by taking the union of the sets of visited models. The sample allocation estimates based on the maximum a posteriori probability and those based on the least-squares clustering

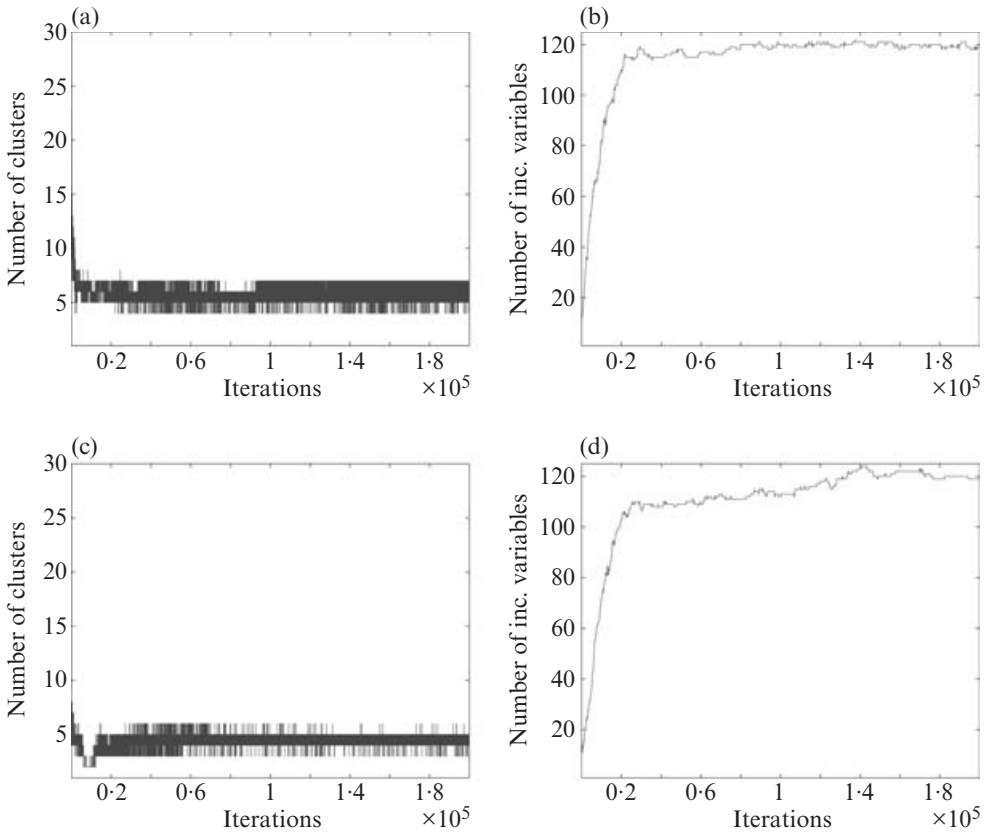

Fig. 3: Microarray data. Trace plots for (a) number of clusters with $\alpha = 38$, (b) number of discriminating variables with $\alpha = 38$, (c) number of clusters with $\alpha = 1$, (d) number of discriminating variables with $\alpha = 1$.

algorithm, as described in § 3, were respectively

$$\hat{c}_{\mathrm{MAP}} = (\underbrace{1, 2, 1, \ldots, 1, 1, 1, \ldots, 1, 1, 1, 1}_{\mathrm{ALL}}, \underbrace{2, 1, 4, 5, 3, 2, 3, 4, 2, 2, 7}_{\mathrm{AML}}),$$

$$\hat{c}_{\mathrm{LSC}} = (\underbrace{1, 2, 1, \ldots, 1, 2, 1, \ldots, 1, 2, 1, 1}_{\mathrm{ALL}}, \underbrace{2, 1, 4, 5, 3, 2, 3, 6, 2, 2, 7}_{\mathrm{AML}}).$$

Figure 4(a) displays a heatmap of the pairwise posterior probabilities, $\mathrm{pr}(c_i = c_j|X)$. The first 27 indices correspond to the acute lymphoblastic leukaemia (ALL) patients and the last 11 to the acute myeloid leukaemia (AML) patients. Except for patient 25, and to a lesser extent patients 2, 12 and 20, all pairs of observations among the ALL group have a high probability of being assigned to the same cluster. The AML group instead exhibits less homogeneity. Thus, all results indicate that we are able to separate successfully the ALL and AML patients and suggest that there may be subgroups among the AML's.

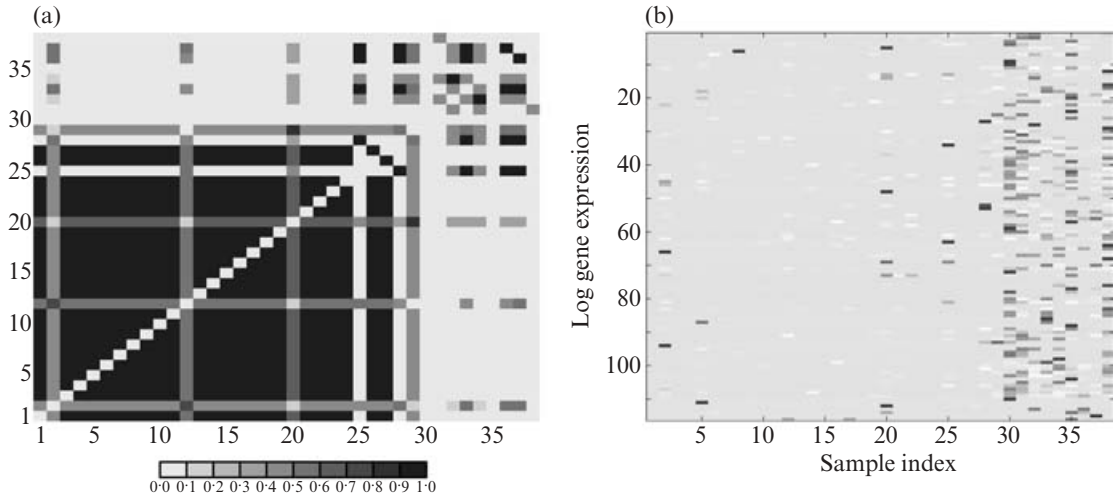

Fig. 4: Microarray data. (a) Pairwise posterior probabilities $\mathrm{pr}(c_i = c_j|X)$, (b) heatmap of selected genes.

For inference on the variable selection, we computed the marginal posterior probabilities of the γ_j's. Differences in marginal posterior probabilities for each gene across the two Markov chain Monte Carlo chains were minimal. There is good concordance in the results despite the different starting points. This suggests that similar regions were visited by the two chains. After pooling the output, we recomputed the marginal posterior probabilities. There were 116 genes with marginal posterior probabilities greater than 0·7. A heatmap of the selected genes is given in Fig. 4(b), where the columns correspond to the samples and the rows represent the log gene expression levels. These genes clearly discriminate the ALL patients, columns 1 to 27, and AML patients, columns 28 to 38. We also looked at the genes selected based on the $\hat{\gamma}$ vector from equation (20). This set contained 120 genes that included all the 116 selected with the marginal inference. A large number of the genes identified by our method as discriminating between the different groups are known to be implicated with the differentiation or progression of leukaemia cells. Some of the selected genes include the Charcot–Leyden crystal protein coding gene, which is known to be down-regulated in AML patients with high white blood cell count, the PRAME gene, which is expressed in acute leukaemia samples, with highest association in AML

tumours carrying t(8; 21) to t(15; 17) chromosomal abnormalities that have a relatively favourable prognosis, and the myeloid cell nuclear differentiation antigen, which is correlated with myeloid and monocytic differentiation of acute leukaemia but is absent in ALL.

We repeated the analysis with $\alpha = 1$. Figures 3(a) and (b) show the corresponding trace plots for the number of clusters and the number of discriminating variables. The sampler visited models with 3 to 6 clusters and around 120 discriminating variables. We note again a slightly slower mixing for smaller values of α, with the chain reaching 120 variables only at around iteration 140 000. The posterior sample allocations were given by

$$\hat{c}_{\mathrm{MAP}} = (\underbrace{1, \ldots, 1, 1, 1, \ldots, 1, 1, 1, 1}_{\mathrm{ALL}}, \underbrace{2, 2, 4, 3, 3, 2, 3, 6, 2, 2, 5}_{\mathrm{AML}}),$$

$$\hat{c}_{\mathrm{LSC}} = (\underbrace{1, \ldots, 1, 2, 1, \ldots, 1, 2, 1, 1}_{\mathrm{ALL}}, \underbrace{2, 2, 4, 5, 3, 2, 3, 6, 2, 2, 5}_{\mathrm{AML}}).$$

The ALL and AML samples are successfully separated. Samples 12 and 25 from the ALL class appear to be closer to some of the observations among the AML group. Again, we note more heterogeneity among the latter, suggesting potential AML subtypes. The posterior inference on the variable selection identified 100 genes based on marginal posterior probabilities greater than 0·7, and 112 genes based on the $\hat{\gamma}$ vector with highest joint posterior probability. These were all included in the set of discriminating genes identified in the previous analysis.

5. Discussion

The use of infinite mixture models is an attractive alternative to finite mixture models, which require a dimension-jumping technique to create and delete clusters. With the Dirichlet process mixture models, the creation and deletion of clusters is naturally taken care of in the process of updating the sample allocations.

We have adopted two approaches for estimating the sample allocations. One could also draw inference conditional on a fixed number of clusters, for instance by conditioning on the value most frequently visited by the sampler. However, this has the limitation of using only a subset of the Markov chain Monte Carlo output. In addition, with the Gibbs sampling update adopted here, a label-switching problem arises since the likelihood is invariant under permutation of the component labels. This problem can be handled using Stephens' relabelling algorithm, in which the Markov chain Monte Carlo output is post-processed to minimise an appropriate loss function (Stephens, 2000b). Alternative posterior estimators can also be obtained by using the Rao–Blackwellisation method or by using decision theoretic approaches.

Acknowledgement

We are grateful to an anonymous referee, the associate editor and the editor for their careful reading of the paper and their constructive comments. We also thank David Dahl and Peter Müller for useful discussions. This research was supported by a grant from the U.S. National Institutes of Health.

References

Abramowitz, M. & Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Washington DC: U.S. National Bureau of Standards.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–74.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–21.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–5.

Brown, P. J. (1993). *Measurement, Regression and Calibration*. Oxford: Clarendon Press.

Brown, P. J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B **60**, 627–41.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics*, Ed. K.-A. Do, P. Müller and M. Vannucci. Cambridge University Press. To appear.

Dudoit, S., Fridlyand, J. & Speed, T. P. (2002). Comparison of discimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Assoc.* **97**, 77–87.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Assoc.* **89**, 268–77.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–88.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, Ed. H. Rizvi and J. Rustagi, pp. 287–302. New York: Academic Press.

Friedman, J. H. & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with Discussion). *J. R. Statist. Soc.* B **66**, 815–49.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339–73.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*, Ed. E. M. Keramigas, pp. 156–63. Fairfax Station, VA: Interface Foundation.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–7.

Green, P. J. & Richardson, S. (2001). Modeling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.* **28**, 355–75.

Hoff, P. D. (2006). Model-based subspace clustering. *Bayesian Anal.* **1**, 321–44.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.

Jain, S. & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comp. Graph. Statist.* **13**, 158–82.

Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Am. Statist. Assoc.* **91**, 1343–70.

Liu, J. S., Zhang, J. L., Palumbo, M. L. & Lawrence, C. E. (2003). Bayesian clustering with variable and transformation selections. In *Bayesian Statistics 7*, Ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, pp. 249–75. Oxford University Press.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist.* B **23**, 727–41.

MacEachern, S. N. & Müller, P. (1998). Estimating mixtures of Dirichlet process models. *J. Comp. Graph. Statist.* **7**, 223–38.

MacEachern, S. N., Clyde, M. & Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Can. J. Statist.* **27**, 251–67.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.* **63**, 215–32.

McLachlan, G. J. & Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

Medvedovic, M. & Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–206.

Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis,*

Seattle, 1991, Ed. C. R. Smith, G. J. Erickson and P. O. Neudorfer, pp. 197–211. Dordrecht: Kluwer Academic Publishers.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Statist.* **9**, 249–65.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C. & Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60**, 812–9.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *Ann. Statist.* **28**, 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *J. R. Statist. Soc.* B **62**, 795–809.

Tadesse, M. G., Sha, N. & Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Am. Statist. Assoc.* **100**, 602–17.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. R. Statist. Soc.* B **62**, 159–80.

[*Received December* 2004. *Revised March* 2006]